FACULTADE DE FILOLOXÍA DA UNIVERSIDADE DA CORUÑA

Departamento de Galego-Portugués, Francés e Lingüística

ALGORITMOS DE PROCESSAMENTO DA LINGUAGEM NATURAL PARA SISTEMAS DE CONVERSÃO TEXTO-FALA EM PORTUGUÊS

Daniela Filipa Macedo Braga Moreira da Silva

Dissertação submetida para obtenção da "mención de DOUTOR EUROPEO"

Dissertação realizada sob a direcção de:

Professor Doutor Xosé Ramón Freixeiro Mato
(Universidade da Coruña)

Professora Doutora Maria Aldina Marques
(Universidade do Minho)

Professor Doutor Fernando Gil Vianna Resende Jr.
(Universidade Federal do Rio de Janeiro)

A Coruña, 23 de Maio de 2008

FACULTADE DE FILOLOXÍA DA UNIVERSIDADE DA CORUÑA

Departamento de Galego-Portugués, Francés e Lingüística

ALGORITMOS DE PROCESSAMENTO DA LINGUAGEM NATURAL PARA SISTEMAS DE CONVERSÃO TEXTO-FALA EM PORTUGUÊS

Dissertação submetida para obtenção da "mención de DOUTOR EUROPEO"

A autora da tese

Vº e praz

JÚRI

PRESIDENTE:
PROF.ª DRA. NIEVES RODRÍGUEZ BRISABOA (DEPARTAMENTO DE COMPUTACIÓN, UNIVERSIDADE DA CORUÑA, ESPAÑA)

VOGAIS:
PROF.ª DRA. M. ANTONIA MARTI ANTONÍN (DEPARTAMENTO DE LINGÜÍSTICA GENERAL, UNIVERSIDAD DE BARCELONA, ESPAÑA)

PROF. DR. JOSÉ JOÃO ALMEIDA (DEPARTAMENTO DE INFORMÁTICA, UNIVERSIDADE DO MINHO, PORTUGAL)

PROF. DR. ANTÓNIO TEIXEIRA (DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA, UNIVERSIDADE DE AVEIRO, PORTUGAL)

SECRETÁRIO:
PROF. DR. ÁLVARO IRIARTE SANROMÁN (DEPARTAMENTO DE ESTUDOS PORTUGUESES, UNIVERSIDADE DO MINHO, PORTUGAL)

JÚRI SUPLENTE
PROF. DR. MANUEL FERREIRO FERNÁNDEZ (DEPARTAMENTO DE GALEGO-PORTUGUÉS, FRANCÉS E LINGÜÍSTICA, UNIVERSIDADE DA CORUÑA, ESPAÑA)
PROF.ª DRA. M. FÁTIMA SILVA (DEPARTAMENTO DE ESTUDOS PORTUGUESES E ESTUDOS ROMÂNICOS, UNIVERSIDADE DO PORTO)

CLASSIFICAÇÃO OBTIDA: "SOBRESALIENTE CUM LAUDE"

Ao Luís.

À minha mãe.

" O Universo está escrito em linguagem matemática."

"É preciso eliminar os mal-entendidos entre a fé e a ciência."

"Quando alguém menos entende mais quer discordar."

"Eu creio na razão."

"Eppur si Muove!"

Galileu Galilei (1564-1642)

Índice

Agradecimentos

Em primeiro lugar, os meus sinceros agradecimentos vão para os meus queridos orientadores. Ao Professor Freixeiro, com quem tanto aprendi sobre Gramática Galega, por todas as palavras de ânimo e por todo o apoio incondicional que sempre teve comigo. Ao Professor Gil, que muito admiro, o verdadeiro catalisador desta tese, agradeço-lhe por tantas horas de reuniões por skype que me dedicou, pela hospitalidade com que me recebeu no Brasil e por toda a amizade que sempre demonstrou. À minha querida e admirável Professora Aldina, ao lado de quem já atravessei o mestrado e com quem continuo a aprender ao longo destes anos. O meu obrigado por toda a força e pelos sábios conselhos que sempre me deu.

Em segundo lugar, as minhas palavras de gratidão seguem para o meu querido esposo, Luís Coelho, por todo o carinho e compreensão, pelos debates científicos que sempre temos, pelos seus bons conselhos e por ser o melhor companheiro de vida do mundo. À minha família, em especial à minha mãe e irmã, o meu obrigado também pelo ânimo que me deram sempre.

Não posso deixar ainda de agradecer ao Professor Miguel Sales Dias, por toda a compreensão que demonstrou sempre que eu necessitei de tempo e de isolamento para me dedicar a este trabalho e aos meus colegas de trabalho, em especial ao Pedro Silva, por me substituir sempre que foi necessário.

E ainda pelos trabalhos conjuntos que temos tido, o meu sincero obrigado ao Ranniery Maia, cujo trabalho muito admiro, e ao Denilson Silva, por toda a simpatia e disponibilidade que sempre demonstrou.

Gostaria ainda de deixar os meus agradecimentos ao Centro Ramón Piñeiro para a Investigación en Humanidades pela cedência dos excertos do CORGA que me permitiram testar os vários módulos apresentados nesta dissertação com textos reais em Galego.

Não posso deixar de retribuir ainda a José Manuel Asorey, director do Servizo de Recursos Audiovisuais da Universidade da Coruña pelo seu apoio e trabalho durante a gravação da base de dados.

Por último, os meus agradecimentos muitos especiais aos Professores José Teixeira e Miguel Sales Dias, que constituíram o meu júri prévio de avaliação da tese, pelas suas sugestões muito pertinentes e enriquecedoras.

Resumo

Em termos gerais, podemos considerar três blocos nos sistemas de conversão texto-fala ou TTS (Text-to-Speech): o *front-end,* que transforma o texto em etiquetas fonéticas através da análise linguística, o *back-end* ou motor de síntese, que converte as etiquetas fonéticas em formas de onda, e a *voice font*, ou base de dados de fala, que é recombinada e treinada pelo *back-end*. A dimensão e qualidade da base de dados dependem da técnica de síntese usada pelo *back-end* e neste ponto o estado da arte apresenta já uma grande maturidade e domínio técnico. Já o *front-end* é dependente das especificidades da língua, e apesar de quase duas décadas de trabalho em português, as soluções ainda se mostram pouco satisfatórias para várias questões relevantes, nomeadamente ao nível da leitura de estrangeirismos, da desambiguação de homógrafos e mesmo da conversão grafema-fone.

Neste trabalho, descreve-se o desenvolvimento dos vários módulos que constituem o pré-processamento ou normalização de texto (separação de frases e de palavras, expansão de abreviaturas, conversão de símbolos e caracteres especiais, conversão de siglas e acrónimos, leitura de numerais árabes cardinais e ordinais, leitura de números romanos, leitura de horas, datas, números com casas decimais, medidas e pontuação desportiva) e a análise fonética (desambiguação de homógrafos, leitura de estrangeirismos, divisão silábica, marcação de sílaba tónica e transcrição grafema-fone) de um sintetizador de fala em português europeu. Os vários módulos foram implementados e testados, tendo sido obtidas as seguintes taxas de acerto: 99,88% para o conversor de siglas/acrónimos e números romanos; 99,86% para o conversor de dígitos (numerais árabes cardinais e ordinais, datas, horas, números com casas decimais, medidas e pontuação desportiva); 98,2% para o desambiguador de homógrafos; 98,14% para o leitor de estrangeirismos; 99,06% para o divisor silábico, 99,54% para o marcador de sílaba tónica e 99,11% para o transcritor grafema-fone.

A grande unidade linguística existente entre o português europeu (PE), o português do Brasil (PB) e a relação histórica entre o português e o galego possibilitam uma elevada adaptabilidade dos algoritmos propostos a essas variedades/línguas. Assim, foi discutida a aplicabilidade de todos os módulos apresentados, o que se traduziu nas seguintes taxas de acerto para o PB: 97,71% para o desambiguador de homógrafos, 99,20% para o divisor silábico e 99,60% para o marcador de tonicidade. Foi ainda desenvolvido um transcritor grafema-fone para PB, adaptado a partir da proposta para PE, que resultou em 97,44% de fones correctamente transcritos. Para o galego, das aplicações directas do divisor silábico e marcador de tónica obtiveram-se 97,87% e 98,52% de acertos para cada um dos módulos testados. Desenvolveu-se ainda um transcritor grafema-fone para galego, seguindo a mesma metodologia, com uma taxa de sucesso de 98,50%.

Seguimos na presente dissertação uma metodologia assente na proposição de regras linguísticas, partindo da consciência de que o português é uma língua flexional, com grande regularidade fonética e fonológica e com uma ortografia de base fonológica. Grande parte deste trabalho foi objecto de publicação em revistas e conferências científicas de âmbito nacional e internacional com revisores.

Palavras-chave: síntese da fala, conversão texto-fala, processamento da linguagem natural, pré-processamento de texto, desambiguação de homógrafos, leitura de estrangeirismos, divisão silábica, marcação de sílaba tónica, conversão grafema-fone, português europeu, português do Brasil, galego.

Abstract

In brief, the architecture of a Text-to-Speech system can be divided in three parts: the front-end, which converts input text into phonetic labels after the linguistic analysis, the back-end or runtime engine, which converts the phonetic labels into waveforms, and the voice font or speech database, which is recombined and trained by the back-end. The size and quality of the voice font depend on the chosen synthesis technique, which is already at a very reliable and mature state of the art. However, the front-end is language-dependent, and in spite of almost two decades of research work in Portuguese speech synthesis, there are still issues to be fully solved, namely regarding the foreign words reading, homograph ambiguity resolution and grapheme-to-phone conversion.

In this work, we describe the development of the modules involved in text normalization (i.e., sentence separator, word breaker, abbreviations' expansion, acronyms and sequences of letters reading, Arabic and Roman numerals conversion and other digits' conversion, such as dates, time, decimal numbers, scores, measurements) and phonetic analysis (homograph ambiguity resolution, foreign words' reading, syllable boundary marking, stress marking and grapheme-to-phone transcription) for a speech synthesizer in European Portuguese. The presented modules were implemented and tested giving rise to the following success rates: 99.88% to the acronyms, sequences of letters and Roman numerals converter; 99.86% for the Arabic numerals converter (ordinal and cardinal numerals, dates, time, percentages, scores, measurements); 98.2% for homograph ambiguity resolution; 98.14% for foreign word reading; 99.06% for the syllable boundary marking, 99.54% for stress marking and 99.11% for grapheme-to-phone transcription.

The linguistic unit between European Portuguese (EP), Brazilian Portuguese (BP) and the historic relationship between Portuguese and Galician explain the high adaptability of the proposed algorithms to those languages/varieties. The algorithms' expansion to BP was discussed, giving rise to the following accuracy rates: 97.71% for homograph ambiguity resolution; 99.20% for syllable boundary marking and 99.60% for stress marking. A grapheme-to-phone transcription module was also developed and adapted to BP. Results of the tests performed to the BP grapheme-to-phone algorithm gave rise to 97.44% of correctly transcribed phones. Direct applications and tests of the syllable boundary marker and stress marker to Galician produced the following accuracy rates: 97.87% for syllable boundary marker and 98.52% for stress marking. A grapheme-to-phone transcriber was also developed to Galician and 98.50% of success rate was obtained.

In this PhD dissertation, a linguistic rule-based methodology was proposed and followed, since Portuguese is an inflectional language, with high phonetic and phonological regularity and with phonologically-based orthography. A great part of this work has been published in international and national journals and conferences with reviewers.

Key-words: speech synthesis, text-to-speech (TTS), natural language processing, text normalization, homograph ambiguity resolution, foreign words reading, syllabification, stress mark, grapheme-to-phone conversion, European Portuguese, Brazilian Portuguese, Galician.

Resumen

En términos generales, se pueden considerar tres partes en los sistemas de conversión texto-voz o TTS (Text-to-Speech): un *front-end*, que transforma el texto en etiquetas fonéticas a través del análisis lingüístico, un *back-end* o motor de síntesis, que convierte las etiquetas fonéticas en formas de onda, y la voice font o base de datos de habla, que es entrenada por el *back-end*. La dimensión y calidad de la base de datos dependen de la técnica de síntesis usada por el *back-end* y, en lo que se refiere a este tema, el estado del arte se presenta ya bastante avanzado. Sin embargo, el *front-end* depende de las especificidades de la lengua, y a pesar de casi dos décadas de trabajo en síntesis del portugués, las soluciones planteadas para varias cuestiones son insuficientes, particularmente con relación a la lectura de extranjerismos, a la resolución de ambigüedad en homógrafos y incluso a la conversión grafema-fone.

En este trabajo, se describe el desarrollo de varios módulos que constituyen el preprocesado o normalización de texto (separación de frases y de palabras, expansión de abreviaturas, conversión de símbolos y caracteres especiales, conversión de siglas y acrónimos, lectura de numerales árabes cardinales y ordinales, lectura de números romanos, lectura de horas, fechas, números decimales, medidas y puntuación deportiva) y el análisis fonético (resolución de ambigüedad de homógrafos, lectura de extranjerismos, separación silábica, marcación de sílaba tónica y conversión grafema-fone) de un sintetizador de habla en portugués europeo. Los varios módulos fueron implementados y como resultados de los testes se han obtenido: 99,88% para el conversor de siglas/acrónimos y números romanos; 99,86% para el conversor de numerales (árabes cardinales y ordinales, fechas, horas, números decimales, medidas y puntuación deportiva); 98,2% para la resolución de ambigüedad de homógrafos; 98,14% para la lectura de extranjerismos; 99,06% para la separación silábica, 99,54% para la marcación de sílaba tónica y 99,11% para la transcripción grafema-fone.

La grande unidad lingüística existente entre el portugués europeo (PE), el portugués de Brasil (PB) y la relación histórica entre el portugués y el gallego hacen posible que haya una grande adaptabilidad de los algoritmos propuestos a esas variedades/lenguas. Por eso, se discutió la aplicabilidad de los módulos presentados con lo que se obtuvieron las siguientes tasas de éxito para PB: 97,71% para la resolución de ambigüedad de homógrafos, 99,20% para la separación silábica y 99,60% para la marcación de sílaba tónica. Se ha desarrollado también un transcriptor grafema-fone para PB, adaptado a partir de la propuesta planteada para PE, com lo que se obtuvo una tasa de 97,44% de fones correctamente transcritos. Aplicaciones directas hechas al gallego de los módulos de separación silábica y marcación de sílaba tónica ofrecen una tasa de éxito de 97,87% y 98,52% respectivamente. Se ha incluso

desarrollado un transcriptor grafema-fone para gallego, siguiendo la misma metodología, con la que se obtuvo 98,50% de tasa de acierto.
En esta disertación doctoral se ha seguido una metodología basada en la proposición de reglas lingüísticas, partiendo del presupuesto de que el portugués es una lengua flexional, con gran regularidad fonética y fonológica y con una ortografía de base fonológica. Gran parte de este trabajo ha sido publicada en revistas y congresos científicos de ámbito nacional e internacional con revisores.

Palabras-clave: síntesis del habla, conversión texto-voz, procesamiento del lenguaje natural, normalización de texto, resolución de ambigüedad de homógrafos, lectura de extranjerismos, separación silábica, marcación de sílaba tónica, conversión grafema-fone, portugués europeo, portugués de Brasil, gallego.

Lista de Tabelas

Lista de Figuras

Lista de Siglas e Abreviaturas

BC – Biblioteca de combinatória lexical restrita

G2P – Grapheme-to-Phoneme

HMMs – Hidden Markov Models

LPS – Laboratório de Processamento de Sinais

LTS – Letter-to-Sound

PB – Português do Brasil

PE – Português europeu

PER – Phone error rate

POS – Part-of-Speech

TTS – Text-to-Speech

UFRJ - Universidade Federal do Rio de Janeiro

WER – Word error rate

WN – Wordnet

UDC – Universidade da Coruña

UM – Universidade do Minho

Introdução

Antecedentes e motivações

> (…) la revolución tecnológica que ha significado la aparición de los nuevos sistemas de comunicación, está provocando un cambio en profundidad en las profesiones y sectores productivos relacionados con el lenguaje, como puedan ser la traducción, la edición – en especial de diccionarios y enciclopedias – y la enseñanza.
> (Listerri & Martí, 2002: 14)

> Whenever I fire a linguist, our system performance improves.
> (Frederick Jelinek, 1988)[1]

Há meio século atrás, pensar-se em máquinas falantes e/ou máquinas ouvintes seria talvez um exercício difícil de imaginação, só possível no domínio da ficção científica. Em 1968, o filme *2001: Odisseia no Espaço* colocou nas telas o futurista computador *Hal*, numa clara previsão dos cenários que a interacção homem-máquina viria a ter. Na década de 80, na série policial *Knight Rider*, *Kitt*, um carro inteligente e falante acompanhava o herói Michael Knight (David Hasselhoff) nas suas investigações. Volvido o ano 2001, ainda parece que estamos longe da autonomia e naturalidade demonstradas pelo *Hal* ou pelo *Kitt*, mas a ficção tornou-se realidade: as línguas conquistaram novos utentes, as máquinas.

O nosso mundo mudou radicalmente nas últimas décadas devido à explosão das novas tecnologias e ao aparecimento da Internet, o que criou novos paradigmas de acesso à informação e ao conhecimento. A informação está agora ao alcance de (quase) todos, através de uma rede gigantesca. Por esta razão, a navegação pelos oceanos da informação revela-se ao mesmo tempo difícil, necessitando de mecanismos que a tornem rápida, eficaz, simples e flexível. Esta espantosa sociedade de informação em que vivemos, sob o signo da revolução tecnológica, não pode, portanto, ser concebida sem a língua nem sem as tecnologias linguísticas, que suportam o diálogo entre o homem e as máquinas e que possibilitam o processamento da informação.

---

[1] Citação muito famosa na área do Processamento da Linguagem Natural e do Processamento da Fala, mas que ao que parece nunca foi escrita. Terá sido usada por Jelinek em 1988, à data membro do *Speech Group* da IBM, na sua comunicação "Applying Information Theoretic Methods: Evaluation of Grammar Quality", in *Workshop on Evaluation of NLP Systems*, Wayne PA (trabalho não publicado).

O processamento da fala[2] em particular e o processamento da linguagem natural[3] em geral constituem, nos dias de hoje, o núcleo das tecnologias linguísticas e um dos novos paradigmas da Língua e da Linguística. À semelhança dos humanos, é necessário ensinar as máquinas a falar e a escrever, porque as máquinas realmente facilitam e agilizam todas as tarefas do nosso quotidiano a um ponto tal que se tornam indispensáveis. Já ninguém imagina a sua vida sem automóvel, telemóvel ou computador. Já ninguém imagina sequer a sua vida sem Internet. Todavia, e por contraste com os humanos, as máquinas não fazem aquisições desestruturadas do mundo para depois as reunir num constructo coerente. As máquinas não fazem extrapolações, induções, deduções ou pressuposições. É necessário treiná-las, programá-las, simular as redes neuronais do nosso cérebro, gerar artificialmente os modelos de processamento da linguagem e do conhecimento que possuímos. Eis o desafio desta tarefa de ensinar as máquinas.

As tecnologias da fala destinam-se a facilitar a interacção entre o utilizador e as máquinas (Mariño *et al*., 1987; Listerri & Martí, 2002: 20), por exemplo, complementando ou substituindo o teclado e o rato. Tradicionalmente, são duas as áreas incluídas no Processamento da Fala: a Síntese da Fala, sistema que permite a conversão de texto em fala, e o reconhecimento de voz, sistema que possibilita a conversão de voz em texto. São inúmeras as vantagens associadas às tecnologias da fala. A primeira de todas é a facilidade de aquisição de informação ao mesmo tempo que é possível realizar tarefas simultâneas, uma vez que liberta os olhos e as mãos do utilizador:

> El uso de la lengua oral como modo de comunicación con los sistemas informáticos libera las manos y la vista, facilitando la recepción de información desde cualquier punto y haciendo posible la movilidad; al mismo tiempo, permite efectuar otras actividades simultáneas (Listerri & Martí, 2002: 20).

Exemplos desta vantagem podem ser observados em sistemas de gestão de stocks de armazéns controlados por voz, possibilitando aos utilizadores realizar tarefas simultâneas de supervisão dos armazéns, no acesso a várias informações pelos pilotos de aviões enquanto fazem o controlo da aeronave, ou no controlo do rádio, do GPS, dos vidros eléctricos ou do ar condicionado dentro do automóvel.

---

[2] O Processamento da Fala é uma sub-área do Processamento da Linguagem Natural que se dedica especificamente ao estudo dos sinais de fala e aos métodos de processamento desses sinais. São alguns dos principais sub-temas do Processamento da Fala a Síntese da Fala, o Reconhecimento de Voz ou da Fala, o Reconhecimento do Locutor, a Codificação da Fala, o *Speech Enhancement*, a Fonética, a Fonologia, a Prosódia, a Produção e Percepção da Fala, a Fisiologia e Patologias da Voz, os Sistemas de Diálogo, etc.

[3] Segundo Diana Santos (2001), o "Processamento da Linguagem Natural (PLN, ou tratamento das línguas por computador) é uma disciplina que (…) se define como a utilização de conhecimentos sobre a língua e a comunicação humana, tanto para a comunicação com sistemas computacionais como para melhorar a comunicação entre os seres humanos." Esta disciplina engloba o Processamento da Fala. São exemplos de aplicações de PLN, os correctores ortográficos, formatadores e hifenizadores, sistemas de extracção de informação, sumarizadores, tradutores automáticos e semiautomáticos, sistemas de pergunta-resposta, os sistemas interactivos, os conversores texto-fala, os reconhecedores de voz, entre outros.

A segunda grande vantagem é a possibilidade de associação destes sistemas a telemóveis e PDAs (*Personal Digital Assistant*), o que possibilita a mobilidade, tão importante nos dias de hoje. Algumas aplicações neste âmbito são a consulta das páginas amarelas, de sítios com informação meteorológica, de sítios bancários ou sítios com informações sobre horários de comboios ou viagens aéreas.

A terceira vantagem prende-se com a adaptação das tecnologias da fala a aplicações ou produtos que possibilitam a acessibilidade à informação por parte de pessoas com deficiências visuais, pessoas com necessidades especiais ou mesmo idosos e crianças. São bem conhecidos os sistemas de conversão texto-fala integrados em e-books, livros *daisy*[4] e leitores de ecrã para cegos[5]. Os reconhecedores de voz permitem, por exemplo, que deficientes auditivos acedam à informação ao converter a voz em texto. Muitos idosos perdem capacidades físicas ou sofrem afasias, na sequência de acidentes vasculares cerebrais. Também nestes casos as tecnologias da fala podem simular a fala humana e assegurar assim aos idosos a sua capacidade comunicativa.

As áreas de aplicação dos sistemas com interface de voz no nosso quotidiano são quase ilimitadas, encontrando talvez limites apenas pela imaginação. De entre as várias áreas de aplicação de sistemas com interfaces de voz, destacámos o ensino (*e-learning* ou ensino à distância, ensino de línguas, aperfeiçoamento de pronúncia), a tradução, a orientação e navegação, os sistemas de consulta de páginas electrónicas, o *e-banking*, os quiosques digitais (sistemas de *e-commerce*), a acessibilidade para pessoas com deficiências, a ajuda em sistemas terapêuticos para pessoas com patologias da fala (detecção e correcção de patologias de fala), auxiliares médicos (sistemas de pré-diagnóstico e monitorização de doentes).[6] De facto, a Síntese de Fala e o Reconhecimento de Voz caminham lado a lado, quer ao nível das metodologias usadas, quer ao nível das aplicações em que podem ser utilizadas.

O Processamento da Fala é, antes de mais, uma ciência no cruzamento de outras ciências, das quais se salientam a Engenharia, a Informática, a Linguística e a Matemática. Este estatuto interdisciplinar está na base de alguma dificuldade de articulação ao nível da metalinguagem, dos posicionamentos teóricos e das opções metodológicas disponíveis. Além disso, esta condição gera inclusive alguma desconfiança por parte de cada uma das áreas envolvidas em relação às outras, o que decorre da inicial dificuldade de comunicação. Frases como a que citámos em epígrafe são um dos mais famosos exemplos da desconfiança e até descredibilização que foi sendo construída pela Engenharia em relação à Linguística, sobretudo a partir dos anos 80.

---

[4] "Trata-se de uma evolução natural do conceito do livro falado, aproveitando os recursos que a era digital nos proporciona. Assim, em vez de cassetes áudio, um livro DAISY é gravado num CD, o que garante à partida uma maior longevidade, melhor qualidade de som, e melhor arrumação. Um só CD DAISY pode conter o equivalente a 10 (DEZ) cassetes áudio! Os áudio-livros DAISY podem ser ouvidos num comum leitor de CDs com MP3". Fonte: Eletrosertec, in http://www.electrosertec.pt/ (06-12-2007).

[5] Empresas como a Electrosertec são especializadas em soluções técnicas para facilitar a acessibilidade à informação a pessoas com problemas de visão ou deficientes visuais.

[6] Para um resumo das principais aplicações, veja-se Coelho *et al.* (2004).

É muito evidente a escassez de trabalhos mais extensos em português na área da Síntese da Fala, sobretudo quando comparados com a miríade de trabalhos feitos para o inglês ou para o francês. É claro que este facto encontra explicação na dimensão e estatuto do inglês e do francês como línguas francas, decorrentes de conhecidas razões históricas, políticas e económicas, o que não acontece com o português ou com o espanhol, muito embora sejam estas as línguas mais faladas do mundo ocidental enquanto línguas maternas e oficiais, depois do inglês. Apesar da investigação na área da Síntese da Fala em português já possuir uma comunidade científica madura e experiente, continua a sofrer de falta de recursos, de investimento económico e de formação. A título de exemplo, podemos verificar que não existe presentemente, nem em Portugal nem no Brasil, nenhum curso (licenciatura ou mestrado) específico nesta área, estando apenas presente em algumas Universidades (Porto, Lisboa, Aveiro, Federal do Rio, entre outras) como disciplinas optativas de cursos de Engenharia ou de Informática. Além disso, e como consequência do reduzido tamanho da comunidade científica portuguesa e brasileira, a maior parte dos trabalhos são redigidos em inglês, permitindo não só uma divulgação internacional da investigação, como também facilitando a publicação nos circuitos da área, que são sobretudo internacionais. Em consequência disso, existe ainda alguma dificuldade na fixação da terminologia científica desta disciplina em língua portuguesa, bem como uma ausência de definição de muitos conceitos.

Outra motivação importante que nos levou à redacção da presente dissertação foi a escassez de trabalhos, tanto no meio académico como no meio comercial, que de facto publiquem o conjunto total de regras e estrutura interna dos algoritmos subjacentes à conversão grafema-fone do português. Além disso, o inventário léxico da língua está em permanente expansão, o que leva à introdução constante de palavras novas, para as quais é preciso prever uma transcrição. No entanto, são muito poucos os autores que publicam os algoritmos subjacentes a esse módulo. Na verdade, trata-se de um tema que está longe de estar esgotado, como se verá pelo estado da arte sobre este tópico, que será feito no Capítulo 5. Da mesma forma que não estão resolvidos nem os problemas resultantes da ambiguidade dos homógrafos heterófonos, nem resultantes da leitura de estrangeirismos. Apesar de estes grupos de palavras existirem em menor número na língua, a sua leitura incorrecta por um sistema de síntese prejudica gravemente a avaliação da inteligibilidade, introduzindo erros inaceitáveis e até passíveis de ambiguidade para o utilizador.

Outra razão que esteve na origem do tema desta dissertação é a actualidade do tema da análise de texto (módulos de pré-processamento e de conversão texto-fala) na arquitectura geral do sistema de síntese da fala. Taylor (2005) delimita-a bem: o léxico de uma língua está em permanente expansão, o que significa que haverá sempre palavras novas que serão introduzidas na língua, sendo para isso necessário um sistema que converta essas novas palavras em transcrições fonéticas:

> Two main justifications are commonly given for the continuing need for a GTP (Grapheme-to-Phoneme) component. Firstly, there will always be genuinely new words ("blairism", "email", "yuppie") created in the course of time. In addition there are many words which may not be new, but were ignored when the system was originally built and have now become common enough to require proper pronunciation (e.g. "bin laden"). In such cases GTP conversion will always be required. Secondly, GTP rules can be used in cases where memory is limited (Taylor, 2005).

Além disso, e apesar de muitas serem as propostas de solução do tema da conversão texto-fala[7], a verdade é que este assunto está longe de estar resolvido para o português, como se pode confirmar pelas ainda significativas taxas de erro publicadas nos artigos da especialidade[8], como também pelos erros de conversão que ainda permanecem nos actuais sistemas comerciais existentes no mercado.

Finalmente, o postulado da grande regularidade fonética e fonológica, bem como uma grande univocidade grafema-fone(ma), subjacente à maior parte das línguas românicas, faz com que a conversão texto-fala em português, tal como em espanhol, seja muito eficaz usando uma metodologia por regras linguísticas, o que não acontece, por exemplo, com línguas como o inglês:

> A eficácia das regras de transcrição depende da regularidade da relação entre a ortografia e a realização oral da língua. Em línguas como a inglesa e a francesa, esta discordância obriga à utilização de grandes léxicos para a sua transcrição automática. Se a palavra não for encontrada no léxico, a aplicação das regras gerais da língua conduzirá com grande probabilidade a uma transcrição incorrecta. No português europeu, no entanto, a ortografia pode ser considerada de base essencialmente fonológica, ou seja, existe uma elevada regularidade entre a ortografia e a fonética (Oliveira, 1996: 40).

Além disso, e tal como defende Taylor na citação acima, uma abordagem por regras é mais interessante em aplicações em que a memória computacional é limitada, como é o caso dos ambientes móveis. A abordagem por regras permite sempre a leitura de uma palavra nova ou de um estrangeirismo que não tenha sido previsto no léxico.

A grande unidade linguística existente entre o português europeu (PE) e o português do Brasil (PB), assim como a relação histórica entre o português e o galego, constituiu a última motivação deste trabalho, possibilitando uma elevada adaptabilidade dos algoritmos propostos a essas variedades/línguas[9]. Entre 2002 e

---

[7] Em Damper (1999) e Taylor (2005) faz-se um resumo das principais abordagens e metodologias.

[8] Dados sobre este assunto serão indicados nas secções de discussão de resultados nos Capítulos 3, 4 e 5.

[9] Existe uma grande discussão e profusa bibliografia defendendo o estatuto do galego como variedade do português. Para uma lista de referências relacionadas,veja-se: Venâncio, F. 2006. Bibliografia da Língua na Galiza, in http://www.agal-gz.org/modules.php?name=News&file=article&sid=3036 (01-01-2008). Destaque ainda para algumas citações de autores portugueses e galegos, defendendo a unidade linguística galego-portuguesa: "Na actualidade, desde o punto de vista estrictamente lingüístico, ás dúas marxes do Miño fálase o mesmo idioma, pois os dialectos miñotos e transmontanos son unha continuación dos falares galegos, cos que comparten trazos comuns que os diferencian dos do centro e sur de Portugal (...)"Fernández Rei, 1990: 17.); "O galego apresenta-se hoje, pois, como um idioma de condição dupla, consoante nos situemos num ponto de vista linguístico ou institucional. Do ponto de vista institucional, o galego, ao adquirir o estatuto de língua oficial, marcou a sua independência, passando a dispor dos seus próprios instrumentos de normalização, estudo e difusão. Do ponto de vista estritamente linguístico, o galego e o português constituem dois grandes grupos de dialectos de uma mesma língua histórica – o Galego-Português (...)" (Ferreira, et al. 1996: 492.)

2006, no âmbito do Projecto ProFala, ao abrigo do projecto bilateral entre o Laboratório de Sinais e Sistemas da Faculdade de Engenharia da Universidade do Porto, de que fizemos parte, e o Laboratório de Processamento de Sinais da Universidade Federal do Rio de Janeiro, nasceram muitas ideias que estão a ser postas em prática nesta dissertação, sendo a principal delas a convergência das estratégias ao nível do pré-processamento de texto para ambas as variedades do português, com manifesto grau de adaptabilidade. Desde então, como teremos oportunidade de referir neste trabalho, foram conduzidos e publicados vários trabalhos que demonstram a grande aplicabilidade dos algoritmos de análise de texto propostos quer ao português do Brasil (Braga *et al.*, 2006a, Braga *et al.*, 2006b) quer ao galego (Braga & Freixeiro, 2006, Braga & Coelho, 2006).

A língua portuguesa, entendida na sua dimensão lusófona e incluindo os países africanos de expressão portuguesa, é falada por 235 milhões e meio de pessoas[10] enquanto língua materna e oficial, o que justifica em grande parte a criação de um Centro de Desenvolvimento da Linguagem em Portugal[11] pela Microsoft, em Novembro de 2005, e o que deveria justificar também um maior investimento científico e económico no plano das tecnologias da fala. É também esse o objectivo subjacente à redacção desta dissertação: contribuir para o avanço da qualidade da fala sintética em português.

Objectivos e metodologia

O presente trabalho tem como objectivo geral construir os módulos de normalização do texto, conversão grafema-fone, desambiguação de homógrafos e transcrição de estrangeirismos necessários a qualquer sistema de síntese da fala em português europeu e testar o seu grau de aplicabilidade em relação ao português do Brasil e ao galego.

Foram propostos os seguintes objectivos específicos para a presente dissertação:

- Enquadramento do tema da presente dissertação na arquitectura de um sistema de síntese da fala;
- Proposição de regras de pré-processamento de texto, nomeadamente:
    - regras de separação de frases e de palavras para PE, PB e galego;
    - regras de conversão de símbolos e expansão de abreviaturas para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;

[10] Dados oficiais publicados pela CIA – Central Intelligence Agency, mais especificamente 235.530.827 pessoas em todos os países de Língua Oficial Portuguesa: (https://www.cia.gov/library/publications/the-world-factbook/geos/tt.html) (11-06-2007)

[11] Microsoft Language Development Center (http://www.microsoft.com/portugal/mldc) (11-06-2007).

    - regras de expansão de siglas e acrónimos para PE, sua implementação, teste e discussão sobre a adaptabilidade ao PB e ao galego;
    - regras de conversão de numerais para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;
- Proposição de regras de desambiguação de homógrafos para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;
- Proposição de regras de leitura de estrangeirismos para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;
- Proposição de regras de divisão silábica para PE, sua implementação, teste e adaptabilidade ao PB e galego;
- Proposição de regras de marcação de sílaba tónica para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;
- Proposição de regras de transcrição grafema-fone para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;
- Aplicação dos algoritmos propostos:
    - Gravação de uma base de dados de fala em português;
    - Integração de todo o sistema com o motor de síntese.

Seguimos na presente dissertação uma metodologia assente na proposição de regras linguísticas, partindo da consciência de que o português é uma língua flexional, com grande regularidade fonética e cuja ortografia é maioritariamente de base fonológica. As principais alternativas a esta metodologia, que são essencialmente a abordagem probabilística e a abordagem *data-driven* (por aprendizagem através de corpora), serão apresentadas no início de cada capítulo, na secção do estado da arte.

O sucesso da aplicação da metodologia por regras a vários módulos da análise de texto em português europeu, português do Brasil e galego tem sido demonstrado nos seguintes trabalhos já publicados:

1. Simões, C.; Calado, A.; Braga, D.; Teixeira, C., Dias, M; "European Portuguese Accent in Non-native English models for ASR systems", in *12th Iberoamerican Congress in Pattern Recognition - CIARP 2007*, Viña del Mar- Valparaíso, Chile, November 2007, pp. 738-747.
2. Braga, D.; Resende Jr.; F. G.; Marques, M. A. 2007. "Leitor de estrangeirismos para sistemas de conversão Texto-Fala em Português Europeu", in *XIII Encontro Nacional da Associação Portuguesa de Linguística*, 1-3 Outubro de 2007, Évora, Portugal.
3. Braga, D.; Coelho, L.; Resende Jr., F.G.V. 2007. "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", in *Proceedings of Interspeech 2007*, 27-31 Agosto de 2007, Antuérpia, Bélgica.

4. Braga, D. & 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.

5. Braga, D.; Resende Jr., F. G. V. 2007. "Módulos de Processamento de Texto Baseados em Regras para Sistemas de Conversão Texto-Fala em Português Europeu", in Lobo, M. & Coutinho, M. A. (Orgs), *XXI Encontro da Associação Portuguesa de Linguística*. Coimbra, 2-4 Outubro de 2006.pp.141-156.

6. Braga, D.; Coelho, L. 2006. "Letter-to-sound conversion for Galician TTS systems", in *IV Jornadas en Tecnologías del Habla*, 8-10 de Novembro de 2006, Zaragoza, Espanha. pp. 171-176. ISBN: 84-96214-82-6.

7. Braga, D.; Freixeiro, X. 2006. "Algoritmos de Conversão Grafema-Fone em Galego para Sistemas de Conversão Texto-Fala", in *VIII Congreso Internacional de Estudos Galegos,* Salvador, Bahía, Brasil, 12-15 de Setembro de 2006 (no prelo).

8. Braga, D.; Coelho, L.; Resende Jr., F. "A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese", in *VI International Telecommunications Symposium (ITS2006)*, 3-6 de Setembro de 2006, Fortaleza-CE, Brasil.

9. Silva, D.; Lima, A.; Maia, R.; Braga, D.; Moraes, J. F.; Moraes, J. A.; Resende Jr. F. "A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing", in *VI International Telecommunications Symposium (ITS2006)*, 3-6 de Setembro de 2006, Fortaleza-CE, Brasil.

10. Braga, D. 2006. "Grapheme-to-phone transcription algorithm for Text-to-Speech Systems in European Portuguese", in *POLISSEMA – Revista de Letras do ISCAP,* vol. nº 6, Instituto Superior de Contabilidade e Administração do Porto, Porto, Portugal.

A metodologia de avaliação e validação dos algoritmos propostos usada ao longo deste trabalho consistiu na:

1. elaboração de casos de teste recolhidos aleatoriamente a partir de corpora reais (bases de dados de textos de vários géneros ou de listas de palavras) e/ou construídos manualmente de forma a prever o maior número de casos possível;
2. construção de uma interface de teste para cada algoritmo;
3. análise dos resultados do output dos algoritmos e comparação com outros resultados descritos na literatura.

Síntese dos conteúdos

Os capítulos centrais, nomeadamente os Capítulos 2, 3, 4 e 5, foram redigidos de forma a conter os seguintes conteúdos: apresentação do problema, estado da arte, arquitectura dos módulos/do sistema, algoritmos de processamento da linguagem natural, teste do sistema, discussão dos resultados, comparação com outras técnicas, conclusões e trabalho futuro. A presente dissertação está estruturada em 7 capítulos, que passamos a resumir:

Capítulo 1. FUNDAMENTOS TEÓRICOS, ESTADO DA ARTE E ARQUITECTURA DO SISTEMA. A natureza híbrida do tema deste trabalho, situado na intersecção entre as áreas da engenharia, da linguística e do processamento da linguagem natural, é explicada neste capítulo. Faz-se uma síntese das principais metodologias utilizadas. Faz-se ainda o elenco dos sistemas de síntese da fala disponíveis para o português em todas as suas variantes, com referência ao estado da arte das tecnologias utilizadas quer no *front-end*, quer no *back-end*. Finalmente, e apesar da grande flutuação existente a nível da arquitectura dos modernos sistemas de síntese da fala, faz-se a descrição da arquitectura do sistema apresentado nesta dissertação, com objectivo de contextualizar a aplicação dos módulos propostos.

Capítulo 2. PRÉ-PROCESSAMENTO DE TEXTO. Também designado por normalização de texto, este é o primeiro módulo a receber o texto que serve de entrada ao sistema de síntese da fala. Neste capítulo, apresenta-se o estado da arte, propõem-se os módulos de separação de frases e de palavras, expansão de siglas e acrónimos, de conversão de numerais e de símbolos, testam-se os vários módulos, discutem-se os resultados e a sua aplicabilidade ao PB e ao galego.

Capítulo 3. DESAMBIGUADOR DE HOMÓGRAFOS. Neste capítulo, após a apresentação do estado da arte relativo a esta questão, apresentou-se uma metodologia inovadora para a desambiguação de homógrafos heterófonos, através de um analisador morfossintáctico ou *POS tagger,* combinado com uma análise semântica, através de bibliotecas de combinatórias lexicais restritas e de bibliotecas de wordnets. Os homógrafos foram divididos em 24 tipos e foi proposto e implementado um algoritmo para cada tipo. Foram conduzidos testes ao sistema e discutidos os seus resultados. Apresentou-se ainda a aplicação dos algoritmos propostos ao PB e ao galego.

Capítulo 4. LEITOR DE ESTRANGEIRISMOS. A leitura de estrangeirismos é outro problema de difícil resolução no actual estado da arte dos sistemas de síntese da fala em português. Neste capítulo propôs-se um módulo de leitura de estrangeirismos, assente na identificação da língua de origem da palavra estrangeira e em conversores grafema-fone para anglicismos e galicismos, por representarem o maior número de palavras estrangeiras na língua portuguesa. O sistema foi implementado e testado. Discutiram-se ainda os seus resultados e a sua aplicabilidade ao PB e ao galego.

Capítulo 5. CONVERSOR GRAFEMA-FONE. Trata-se do capítulo central desta tese. Neste capítulo, apresentou-se o estado da arte sobre este tópico, propuseram-se algoritmos de divisão silábica, de marcação de sílaba tónica e de transcrição grafema-fone para PE, fez-se a sua implementação e teste e discutiram-se os resultados. Apresentou-se ainda a sua adaptação ao PB e ao galego.

Capítulo 6. INTEGRAÇÃO DO SISTEMA NO MOTOR DE SÍNTESE. O objectivo deste capítulo é demonstrar a integração dos módulos de pré-processamento e processamento de texto apresentados previamente num sistema de síntese de fala. Foi para isso gravada uma base de dados ou voice font de 1000 frases em PE que foi anotada foneticamente, através de um sistema de semi-alinhamento automático. A base de dados foi treinada e integrada no motor de síntese segundo a tecnologia do HTS (*HMM-based Text-to-Speech Synthesis*).

Capítulo 7. CONCLUSÕES E TRABALHO FUTURO. Neste capítulo, fez-se uma síntese do trabalho apresentado, bem como uma avaliação dos resultados atingidos em confronto com os objectivos inicialmente propostos. Fez-se ainda uma previsão e análise das principais linhas de acção que este trabalho abre para o futuro.

Capítulo 1

Fundamentos teóricos, estado da arte e arquitectura do sistema

1.1. Fundamentos teóricos

> O conhecimento do paradigma emergente tende assim a ser um conhecimento não dualista, um conhecimento que se funda na superação das distinções tão familiares e óbvias que até há pouco considerávamos insubstituíveis, tais como natureza/cultura, natural/artificial, vivo/inanimado, mente/matéria, observador/observado, subjectivo/objectivo, colectivo/individual, animal/pessoa. Este relativo colapso das distinções dicotómicas repercute-se nas disciplinas científicas que sobre elas se fundaram. (Boaventura de Sousa Santos, 2001: 39-40)

A presente dissertação enquadra-se na área da Linguística Aplicada à Síntese da Fala ou na área do Processamento da Linguagem Natural aplicado à Síntese da Fala. Esta (in)definição de área(s) de conhecimento inscreve-a desde logo no novo paradigma científico e epistemológico anunciado por Boaventura Sousa Santos em 1987, data da 1ª edição do seu "Um discurso sobre as ciências". Sousa Santos descreve o paradigma emergente em quatro postulados[12], sendo o primeiro deles a interdisciplinaridade como nova organização do conhecimento científico. O paradigma científico tradicional assenta na definição clara das fronteiras de cada disciplina, o seu conhecimento é tanto mais rigoroso quanto mais especializado. Nas palavras de Sousa Santos, "é hoje reconhecido que a excessiva parcelização e disciplinarização do saber científico faz do cientista um ignorante especializado". Miguel Martinez (1997: 164), no seu livro "El Paradigma Emergente", explica que esta interdisciplinaridade assenta na busca do conhecimento como um todo e decorre

[12] Sousa Santos descreve o paradigma emergente com as seguintes quatro teses, seguidas de justificação: 1) Todo o conhecimento científico-natural é científico-social; 2) Todo o conhecimento é local e total; 3) Todo o conhecimento é auto-conhecimento e 4) Todo o conhecimento científico visa constituir-se em senso comum. (Sousa Santos, 2001: 37-58).

do desenvolvimento natural das ciências, como fusão de perspectivas separadas. Ainda recentemente, o mesmo autor destaca esse carácter sistémico, relacional e inter/trans-disciplinar do paradigma:

> Es de esperar que el nuevo paradigma emergente sea el que nos permita superar el realismo ingenuo, salir de la asfixia reduccionista y entrar en la lógica de una coherencia integral, sistémica y ecológica, es decir, entrar en una ciencia más universal e integradora, en una ciencia verdaderamente inter- y trans-disciplinaria, como lo propone la UNESCO, donde los diversos puntos de vista, enfoques y abordajes puedan cultivarse a través de un profundo diálogo y ser integrados en un todo coherente y lógico. (Martinez, 2006: 46-47)

É no cruzamento das disciplinas da Linguística, da Informática e da Engenharia que esta dissertação se inscreve, fundindo, por isso, as bases teóricas da Linguística com as metodologias de trabalho e os instrumentos de validação do conhecimento da Engenharia e da Informática.

Ao contrário das teses de mestrado e de doutoramento tradicionais em Linguística, esta dissertação não segue nenhuma corrente teórica, porque na verdade absorve a pragmaticidade funcional de todas elas, que aplica em função de cada problema. Absorve, por exemplo, o princípio estruturalista (independentemente das suas várias actualizações em Glossemática, Funcionalismo, Distribucionalismo ou Generativismo) de que a língua é estrutura e sistema e que, por isso, pode ser descrita por regras. Absorve também o princípio, trazido pelas novas correntes da Linguística do Uso/Funcionamento da Língua (de que se destacam a Análise do Discurso e a Pragmática), de que a língua é também o uso que os falantes fazem dela e que tem de ser estudada, não em função do conhecimento interiorizado e individual do linguista, mas sim à luz de corpora de textos/discursos que permitam fazer deduções generalizadas e, por consequência, fundamentadas. Absorve ainda os conceitos de protótipo e de categorização de um dos mais recentes paradigma da Linguística, o Cognitivismo, uma vez que cada módulo e cada regra apresentados nesta dissertação funcionam como uma forma de organização do nosso conhecimento linguístico. Ao procurarmos simular artificialmente a estrutura e relação entre as partes do conhecimento linguístico humano, partimos assim do conceito cognitivista de categorização como mecanismo de organização e sistematização da informação apreendida na nossa interacção com o mundo.

Além dos fundamentos teóricos da Linguística, a presente dissertação recorre à metodologia científica em que está ancorada a Informática e a Engenharia. A Síntese da Fala é, antes de mais, um ramo das Ciências da Computação, uma vez que se serve das mesmas ferramentas de trabalho, da mesma formulação algorítmica, da mesma linguagem de programação e da mesma metodologia de teste para avaliação e aperfeiçoamento do desempenho. É também um ramo da Engenharia, na medida em que constrói e articula as várias peças de uma engrenagem, de um sistema. Além disso, depende cada vez mais da Linguística para atingir a inteligibilidade e naturalidade da fala sintética, sendo o seu estatuto interdisciplinar bem evidente na necessidade que todos os especialistas e investigadores da área sentem em incluir um capítulo mais ou menos detalhado de Fonética, Fonologia, Prosódia, Sintaxe,

Semântica e Discurso nos seus livros ou teses[13]. Essa interdisciplinaridade sai reforçada se olharmos para os temas que constam dos programas das conferências internacionais da área, como o *Interspeech*[14], já que a Fonética, a Fonologia e a Prosódia, entre outras disciplinas da Linguística, são temas obrigatórios.

1.2. Estado da arte

A óbvia dificuldade inerente a esta tarefa, independentemente da área em análise, associada ao rápido avanço da tecnologia e do conhecimento na área da Síntese da Fala, justificam a nossa opção por apresentar um estado da arte muito genérico, procurando destacar, da melhor forma que nos é possível, o que nos parece ser as principais tendências e os principais actores desta área. De qualquer forma, ao longo de cada capítulo, é feito o estado da arte específico para cada módulo apresentado, o que nos pareceu mais claro do ponto de vista da organização deste trabalho.

Apesar de alguns autores fazerem remontar a história da Síntese da Fala às lendárias "*speaking heads*"[15] medievais, é mais consensual considerarem-se precursores Christian Kratzenstein (1779) e Wolfgang von Kempelen (1791) que construiram modelos do tracto vocal. De qualquer forma, parece indiscutível que a história contemporânea da Síntese da Fala começa nos anos 30 do século XX, quando os Bell Labs lançam aquele que é considerado o primeiro sintetizador comandado por teclado, o Vocoder, que evoluiu para o Voder em 1939. Resumos sobre a história da Síntese da Fala podem ser encontrados em Klatt (1987) e no artigo "Speech Synthesis"[16], da *Wikipedia*, em permanente actualização. Estão também disponíveis na web os arquivos sonoros relativos aos vários sintetizadores que fizeram história e

[13] Destacamos os capítulos dedicados à Fonética, à Análise Morforssintáctica, à Semântica Lexical, ao Discurso e à Tradução Automática da nova edição ainda no prelo de "Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition, de Jurafsky & Martin. Paul Taylor, em outro *draft* que é já uma referência no panorama bibliográfico da área, dedica também um capítulo à Fonética no seu livro "Text-to-Speech Synthesis".

[14] Programa do Interspeech 2007 disponível em: http://www.interspeech2007.org/Calls/cfp.php, em que se destacam os seguintes tópicos da Linguística: *Phonology and phonetics, Discourse and dialogue, Prosody (production, perception, prosodic structure), Paralinguistic and nonlinguistic cues (e.g. emotion and expression), Speech production, Speech perception, Physiology and pathology, Spoken language acquisition, development and learning*.

[15] "A Brazen Head (or Brass Head or Bronze Head) was a prophetic device attributed to many medieval scholars who were believed to be wizards, or who were reputed to be able to answer any question. It was always in the form of a man's head, and it could correctly answer any question asked of it. However, depending on the story, it could be cast in brass or bronze, it could be mechanical or magical, and it could answer freely or it could be restricted to "yes" or "no" answers"(Fonte: http://en.wikipedia.org/wiki/Brazen_Head (06-12-2007).

[16] Disponível em: http://en.wikipedia.org/wiki/Speech_synthesis (06-12-2007).

que compõem os anexos do famoso artigo de Dennis Klatt[17], "Review of text-to-speech conversion for English" (1987).

Pode dizer-se ainda que a história da Síntese da Fala no século XX se faz em função da evolução das técnicas de descodificação do texto em voz. Até ao final dos anos 80, eram as chamadas técnicas de primeira geração (na expressão de Paul Taylor, 2007: 415), nomeadamente a síntese por formantes[18] e a síntese articulatória[19], que dominavam o paradigma dentro desta área. Desde a última década, o estado da arte da Síntese da Fala parece ter estabilizado num bom nível de qualidade com a síntese por concatenação. No novo capítulo sobre *Speech Synthesis*, Jurafski & Martin (*draft* de Outubro de 2007: 36-38)[20] fazem um resumo muito interessante sobre a história das técnicas de síntese e sobre os modelos de sintetizadores mais representativos. Taylor explica que a grande diferença entre as técnicas de primeira geração e as técnicas por concatenação é que estas se constroem a partir de corpora de fala, ao contrário das primeiras, e que efectuam pouca ou até nenhuma modificação desses sinais de fala:

> While details vary we can characterise second generation systems as ones where we directly use data to determine the parameters of the verbal component as with classical linear prediction. The difference is that the source waveform is now also generated in a data driven fashion. The input to the source however is still controlled by an explicit model. So for instance we might have a explicit F0 generation model of the type described in Section 9.5 which generates an F0 value every 10ms. The second generation technique then realises these values in the synthesized speech by a data driven technique, rather than the impulse/noise model (Section13.3.2). The differences in the second generation techniques mainly arise from how explicitly they use a parametrisation of the

[17] Ver em http://www.cs.indiana.edu/rhythmsp/ASA/Contents.html ou em http://cslu.cse.ogi.edu/tts/research/history/ (06-12-2007).

[18] A síntese por formantes, também chamada síntese por regras ("synthesis-by-rule") é uma técnica de primeira geração que consiste num processo de reconstrução de uma onda através da manipulação de certos parâmetros como o *pitch*, os formantes, as larguras de banda dos formantes, etc.: "Formant synthesis adopts a modular, model-based, acoustic-phonetic approach to the synthesis problem. The formant synthesiser makes use of the acoustic tube model, but does so in a particular way so that the control elements of the tube are easily related to acoustic-phonetic properties than can easily be observed." (Taylor, 2007: 398).

[19] A síntese articulatória baseia-se na simulação da produção da voz ao longo do tracto vocal, imitando quer os movimentos mecânicos dos órgãos articuladores quer o efeito da pressão do ar à saída dos pulmões, laringe e cavidades oral e nasal. Nas palavras de Taylor: "The attractive part of articulatory synthesis is that as the tubes themselves are the controls, this is a much easier and more natural way to generate speech; small, "natural" movements in tubes can give rise to the complex patterns of speech, thus bypassing the problems of modelling complex formant trajectories explicitly. Often articulatory synthesis models have an interim stage, where the motion of the tubes is controlled by some simple process (such as mechanical damping, or filtering) intended to model the fact that the articulators move with a certain inherent speed" (Taylor, 2007: 440).

[20] Segunda edição do livro ainda não publicado disponível em: http://www.cs.colorado.edu/~martin/slp2.html.

> signal. While all use a data driven approach, some use an explicit speech model (for example using linear prediction coefficients to model the vocal tract) while others perform little or no modelling at all, and just use "raw" waveforms as the data. (Taylor, 2007: 422-423)

A grande qualidade conseguida pela técnica por concatenação justifica o *boom* das indústrias da fala desde os anos 80, que surgem também como resposta a uma procura crescente de soluções para contextos de *hands-free* e *eyes-free* por parte de um mercado cada vez mais exigente. São já várias as empresas que se dedicam ao desenvolvimento de tecnologias de síntese da fala com alta qualidade, como a italiana Loquendo[21], a belga Acapela[22], a catalã Verbio Speech Technologies[23] (anterior Atlas), ou as multinacionais Siemens[24], Nokia[25], France Telecom[26], Nuance[27] (anterior Scansoft), IBM[28] ou Microsoft[29]. São ainda de destacar as empresas americanas Cepstral[30], At&T[31], Aculab[32] e Fonix[33]. Cada empresa adopta uma estratégia diferente para se afirmar no mercado. Lauren Shopp, na *Speech Technology Magazine*, edição de Novembro/Dezembro de 2007, destaca duas estratégias seguidas pelas empresas que desenvolvem sistemas conversão texto-fala e que resume no sub-título do artigo[34]: "Hundreds of choices or one perfect voice?".

> Two camps have formed within the industry: One pushes for a singular, perfect speaker, while the other seeks variety in the form of 1,000 voices. (Shopp, 2007)

Na verdade, as tendências da indústria parecem ser estas: há empresas que oferecem poucas línguas mas várias vozes disponíveis para cada língua (como é o caso da Cepstral, da IBM, da Acapela e da AT&T), enquanto outras companhias apostam na diversidade de línguas mas apenas uma voz por língua (como por exemplo a Nuance ou a Microsoft). Outra estratégia de afirmação no mercado por

---

[21] Demos disponíveis em: http://www.loquendo.com/en/demos/demo_emb_tts.htm/ (06-12-2007).
[22] Disponível em: http://www.acapela-group.com/ (06-12-2007).
[23] Demos disponíveis em: http://www.verbio.com/webverbio2/html/demos_ttsonline.php (6 06-12-2007).
[24] Disponível em: https://www.ct.siemens.com/en/business/speech/index.html (06-12-2007).
[25] Disponível em: http://www.nokia.com/ (06-12-2007).
[26] Disponível em: http://www.francetelecom.com/fr/groupe/rd/ (06-12-2007).
[27] Disponível em: http://www.nuance.com/ (06-12-2007).
[28] Demos disponíveis em: http://www-306.ibm.com/software/pervasive/voice_server/demos/?S_CMP=rnav (06-12-2007).
[29] Demos disponíveis em: https://research.microsoft.com/speech/tts.asp (06-12-2007).
[30] Disponível em: http://cepstral.com/ (06-12-2007).
[31] Demos disponíveis em: http://www.research.att.com/~ttsweb/tts/demo.php (06-12-2007).
[32] Disponível em http://www.mcct.com/aculab_text.html (06-12-2007).
[33] Disponível em: http://www.fonixspeech.com/tts.php (06-12-2007).
[34] Este artigo está disponível em: http://www.speechtechmag.com/Articles/Editorial/Feature/TTS-Is-Finding-Its-Way-40067.aspx (06-12-2007)

parte da indústria de tecnologias de fala passa pela publicação, que pode ser de dois tipos: académica, reflectindo um perfil voltado para a investigação e mostrando actualização que acompanham o estado da arte científico (Mao *et al.*, 2007[35]; Sündermann *et al.*, 2006[36]; Pollet & Coorman, 2004[37]; veja-se ainda a lista de publicações dos vários grupos de speech da Microsoft Research[38]) e comerciais, sob a forma de *white papers*, documentos técnicos dirigidos ao cliente final, em que se apresentam novas tecnologias e aplicações. Estes *white papers* funcionam, em primeira instância, como estratégia de *marketing*, na medida em que têm como objectivo promover a empresa que os publica, mas o prestígio e qualidade de muitos deles confere autoridade tecnológica a quem os publica. Só para referir alguns exemplos interessantes, vejam-se os artigos da Nuance (Scansoft, 2004, "Assessing Text-to-Speech System Quality")[39] e da Loquendo (Baggia *et al.* 2006; Baggia & Mosso, 2005)[40]. Lemmety (1999: 64-78)[41], na sua dissertação de mestrado, faz também o elenco e descrição dos sintetizadores comerciais disponíveis no mercado até ao momento.

Mas continua a ser no plano académico que se fazem os maiores progressos científicos na área da Síntese da Fala. Actualmente, pode considerar-se que os grandes centros de desenvolvimento científico se concentram na Europa, nos Estados Unidos e na Ásia, especialmente no Japão e na China. Não sendo nosso objectivo elencar todos os centros de investigação sob pena de esquecer muitos deles, mencionaremos alguns que têm trazido um contributo relevante e, em nosso entender, até histórico para a área. No sítio da ISCA[42], uma das principais organizações internacionais dedicadas ao Processamento da Linguagem e da Fala, encontra-se uma lista, ainda que incompleta, de laboratórios dedicados ao Processamento da Fala organizada por país[43]. A mesma organização promove SIGs (Special Interest Groups) em várias áreas, nomeadamente na área da Síntese da Fala[44], onde também se podem

---

[35] Os primeiros autores pertencem à France Telecom R&D Center, em Pequim.

[36] Os primeiros autores pertencem à Siemens. Artigo publicado no ICASSP 2007.

[37] Artigo da Scansoft, actualmente Nuance, publicado no Interspeech 2004.

[38] A Microsoft Research é uma unidade dentro da Microsoft dedicada apenas à investigação. Há dois grupos importantes, o China Speech Group (http://research.microsoft.com/speech/) e o Speech Research Group, em Redmond, nos Estados Unidos (http://research.microsoft.com/srg/). É vasta a sua lista de publicações disponível online, para não falar de um dos poucos livros de referência disponíveis publicados pela Microsoft Research: Huang et al. (2001), "Spoken Language Processing".

[39] Os actuais white papers da Nuance ainda estão com a autoria da Scansoft. É interessante que estes white papers não apresentam autoria, ao contrário dos da Loquendo, e é preciso fazer uma subscrição electrónica para os obter.

[40] Todos os white papers da Loquendo estão disponíveis online em: http://www.loquendo.com/en/company/whitepapers.htm.

[41] Disponível online em: http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap9.html

[42] ISCA – International Speech Communication Association: http://www.isca-speech.org/.

[43] Veja-se em: http://www.isca-speech.org/speech_labs.html.

[44] SynSIG – Speech Synthesis Special Interest Group: http://www.synsig.org/index.php/Main_Page (08-12-2007)

encontrar alguns centros muito activos nesta área. Entretanto, gostaríamos de destacar:

- Na Europa:
  - *Institute of Phonetic Sciences, University of Amsterdam*, Holanda, onde o Praat[45] foi desenvolvido;
  - *KTH – The Royal Institute of Technology*, Suécia, activo na área da Fonologia, onde Gunnar Fant, um dos nomes históricos da Síntese da Fala e da Fonologia, é Professor Emérito; activo ainda na área da Síntese Multi-modal[46];
  - *Department of Phonetics and Linguistics, University Colledge of London*, Reino Unido, autores do SFS[47];
  - *Laboratoire de Parole et Langage, Université de Provence*, laboratório activo na área da Prosódia, dirigido por Daniel Hirts, autor do INTSINT[48];
  - *University of Edimburgh*, Reino Unido, onde Alan Black e Paul Taylor desenvolveram o Festival[49];

- Nos Estados Unidos da América:
  - *Language Technologies Institute, Conergie Mellon University*, onde se desenvolve o projecto Festvox[50], responsável por iniciativas de avaliação de sistemas de síntese da fala, os *Blizzard Challenges*[51], entre outras;
  - *University of Colorado, at Boulder* (Daniel Jurafsky é um dos grandes nomes)
  - *MIT - Massachusetts Institute of Technology*, Estados Unidos, onde se desenvolveu o DAVO, um dos primeiros sintetizadores articulatórios (1958) e o MITTalk, em 1979, um dos primeiros sintetizadores baseados em dicionários (Allen *et al.*, 1987) e o Klattalk, em 1983;

- Na Ásia:

---

[45] O Praat é uma das ferramenta freeware de análise de sinal mais utilizadas pela comunidade científica. Disponível em: http://www.fon.hum.uva.nl/praat/ (06-12-2007).

[46] Vejam-se os projectos e publicações do grupo de Multimodal speech technology: http://www.speech.kth.se/multimodal/ (06-12-2007).

[47] SFS – Speech Filling System, outra ferramenta freeware que compete com o Praat. Disponível em: http://www.phon.ucl.ac.uk/resource/sfs/(06-12-2007).

[48] INTernational Transcription System for INTonation. Para mais informações, ver: http://aune.lpl.univ-aix.fr/~hirst/intsint.html (06-12-2007).

[49] Tookit para gerar sintetizadores; mais informação em: http://www.cstr.ed.ac.uk/projects/festival/(06-12-2007)

[50] Mais informações em: http://www.festvox.org/(06-12-2007).

[51] Ultimos relatórios e informações sobre as avaliações em: http://festvox.org/blizzard/(06-12-2007).

- *Nagoya Institute of Technology*, onde Keiichi Tokuda e o seu grupo desenvolveram o HTS[52], considerado o novo paradigma nas técnicas de síntese;
- ATR[53] – *Advanced Telecommunications Research Institute International*, Japão, um dos maiores centros de desenvolvimento estatais do Japão;
- *Microsoft Research Asia – Speech Technology Group*[54], em Pequim, China, onde tecnologias de ponta estão a ser desenvolvidas.

Outra iniciativa que tem contribuído largamente para o progresso científico é a formação de consórcios e de projectos internacionais, com destaque para os projectos financiados pela União Europeia, de onde resultam conferências e trabalhos conjuntos entre grupos académicos e empresariais com experiências distintas. Neste plano, são assim de ressaltar:

- as acções COST[55], particularmente o COST 258[56] em "The naturalness of synthetic speech", de onde resultou uma publicação relevante (Keller *et al.* 2002);

- o projecto europeu TC-Star - Technology and Corpora for Speech to Speech Translation, financiado pelo Sexto Programa Quadro, que congrega academia e indústria, do qual resultaram relatórios com especificações e linhas de acção para produzir e avaliar os vários módulos que compõem os sistemas de síntese da fala segundo os melhores padrões de qualidade (Bonafonte *et al.*, 2004)[57];

- o *Companions Consortium*[58], também financiado pelo Sexto Programa Quadro, é um consórcio europeu, que reúne empresas e universidades, criado com o objectivo de desenvolver as relações Homem-Máquina, tem como uma das áreas o desenvolvimento de tecnologias de síntese e reconhecimento de fala;

---

[52] HTS é o mesmo que "HMM-Based Speech Synthesis System" (para mais informações consultar: http://hts.sp.nitech.ac.jp/) (06-12-2007).
[53] Disponível em: http://www.atr.jp/index_e.html (06-12-2007).
[54] Website disponível em: http://research.microsoft.com/speech/ (06-12-2007).
[55] Presentemente não está em curso nenhuma acção COST no domínio da Síntese da Fala.
[56] Informação disponível em: http://www.cost.esf.org/index.php?id=110&action_number=258 (06-12-2007).
[57] É de destacar o documento intitulado "TTS baselines and Specifications" , (Bonafonte et al. 2004), disponível em: http://www.tc-star.org/ (06-12-2007).
[58] Mais informações em: http://companions-project.org/(06-12-2007).

- o ECESS – European Center of Excellence in Speech Synthesis[59], um grupo em grande parte saído do TC-Star, que dá continuidade às actividades iniciadas no contexto do TC-Star, incentivando a avaliação e partilha de recursos e sistemas, no sentido de promover o progresso e a excelência na área da Síntese da Fala.

O avanço científico-tecnológico na área das Tecnologias de Fala em geral e da Síntese da Fala em particular faz-se de forma tão rápida que publicações com dez anos se tornam frequentemente ultrapassadas. Este é um aspecto que distingue desde logo a área da Síntese da Fala da área da Linguística, por exemplo. Este aspecto explica que existam poucos livros de carácter propedêutico na área da Síntese e explica também que os seus autores se vejam obrigados a trabalhar nas edições seguintes desde o momento em que publicam o livro. Por esta razão, é fácil listar os livros mais relevantes na área da Síntese, nomeadamente: o clássico Rabiner (1978), e os mais recentes Deller *et al.* (2000), Dutoit (2001), Huang *et al.* (2001), Taylor (2008, no prelo), Jurafsky & Martin (2007 – 2ª edição no prelo). Destaque ainda para o tutorial online de Alan Black, disponível em: http://festvox.org/festtut/notes/festtut_toc.html.

O estado da arte na área da Síntese da Fala altera-se significativamente a cada ano que passa, podendo em grande parte ser acompanhado através das publicações em congressos internacionais de periodicidade anual como o Interspeech e o ICASSP, em workshops de Speech Synthesis promovidas pela ISCA[60], ou em revistas da especialidade como *Computer Speech and Language*, *IEEE Transactions on Audio, Speech, and Language Processing* e *The ACM Transactions*.

Quanto aos grandes desafios que dominam neste momento o estado da arte em Síntese da Fala, partiremos do artigo que Gerard Bailly, Nick Campbell e Bernd Mobius publicaram no Interspeech 2003. Neste artigo, intitulado "ISCA Special Session: Hot topics in Speech Synthesis", começa-se por uma análise dos grandes temas que dominavam o estado da arte em Síntese da Fala no ano de 2003 e termina-se com uma previsão de quais serão as suas possíveis áreas de aplicação em 2008. Apesar do rápido avanço tecnológico, e encontrando-nos nós no limiar do ano de 2008, é possível concluir que, mais do que a concretização das suas previsões, ainda se mantêm muito actuais os grandes temas do ano de 2003 elencados pelos autores, designadamente:

1. **Avaliação**; este tópico foi apresentado pelos autores em primeiro lugar da lista, mas mantém-se ainda na transição entre 2007 e 2008 como um dos grandes temas da actualidade em Síntese da Fala; veja-se a realização dos Blizzard Challenges[61] em 2005 (Black & Tokuda, 2005), 2006 (Bennet & Black, 2006) e 2007 (Fraser & King, 2007), com uma

[59]Mais informações em: http://www.ecess.eu/ (06-12-2007).
[60] ISCA – International Speech Communication Association: http://www.isca-speech.org/.
[61] Anúncio do Blizzard Challenge 2008, resultados e papers de Blizzards anteriores disponíveis em: http://festvox.org/blizzard/ (08-12-2007).

adesão cada vez maior de sistemas a concurso[62]e veja-se os sucessivos relatórios de avaliação de sistemas comerciais disponíveis publicados pela ASR News[63]; destaque ainda para as actividades do ECESS, dedicadas à avaliação de ferramentas, sistemas, recursos linguísticos e módulos de TTS. Para um resumo dos vários tipos de testes para avaliação da qualidade, inteligibilidade e naturalidade da fala sintética, veja-se ainda Lemmety, (1999, capítulo 10).

2. **Síntese multi-língua**; este é ainda um tópico muito presente nos programas das conferências internacionais da área; o conceito de possuir um aparelho que contenha sintetizadores em várias línguas não é novo (Black & Lenzo, 2004; Shalonova & Tucker, 2003; projecto EULER[64], 2000; vejam-se as demonstrações da síntese multi-língua na página pessoa de Tokuda[65]); a dificuldade reside na partilha dos vários módulos que compõem os vários sintetizadores.

3. **Emoção/Expressividade**; este tópico, apresentado pelos autores em terceiro lugar, deveria passar para segundo lugar na actual lista de assuntos quentes e ainda longe de estarem resolvidos; a prosódia, as emoções e mais do que isso, a expressividade (atitude e estilo) representam o último passo para a naturalidade da fala sintética; são inúmeros os trabalhos dedicados a este assunto (Cahn, 1990; Bulutl *et al.*, 2002; Hamza *et al.*, 2004; Eide *et al.*, 2004; Lee *et al.*, 2006; Schroder, 2006; veja-se ainda, como referência, os trabalhos publicados nos congressos internacionais Interspeech e Speech Prosody); algumas empresas inclusive já comercializam sistemas de síntese com emoções, como a Loquendo e a Cereproc[66].

---

[62] "The Blizzard Challenge 2005 [1, 3] had 6 participants and Blizzard 2006 had 14 [4]. In 2007, the number of entries increased again: 19 sítios registered, 18 returned signed licences for the data and 16 submitted entries" (Fraser & King, 2007).

[63] Vejam-se os resultados de "Text-to-Speech Accuracy Testing - 2005" e "Text-to-Speech Accuracy Testing – 2006" disponíveis em: http://www.asrnews.com/accuracy.htm (08-12-2007).

[64] Resumo do projecto: "EULER is a collaborative R&D project set up by the Speech synthesis research group of the Faculté Poytechnique de Mons. Its objective is to provide a freely available, easy-to-use, and easy-to-extend, generic multilingual TTS for Windows95/NT, Mac-OS, and UNIX which will progressively integrate results of existing multilingual language and speech processing projects", disponível em http://tcts.fpms.ac.be/synthesis/euler/ (08-12-2007).

[65] Disponível em: http://www.sp.nitech.ac.jp/~tokuda/HTS_demo/multilingual/index.html (08-12-2007).

[66] A Cereproc, sediada em Edimburgo, comercializa sistemas de síntese para o inglês e suas variantes a que junta emoções. Veja-se mais informações em: http://www.cereproc.com/rand.html (10-12-2007).

4. **Síntese multi-modal ou síntese audio-visual** é a combinação de voz, gestos, movimentos dos olhos e expressões faciais; como destacam os autores, a síntese multi-modal prova que a fala acompanhada de informação visual possibilita uma comunicação mais eficaz, particularmente em ambientes com ruído. Sobre o assunto, vejam-se Beskow (2003)[67] Waibel *et al.* (2007)[68] e Theobald (2007). Vejam-se ainda as "talking heads" do grupo de tecnologia multimodal do KTH[69].

5. **Inputs e sistemas de mark-up** são assuntos de não menor importância, como referem Bailly et al (2003). O VXML (Voice XML) será muito necessário à medida que os sintetizadores evoluírem no sentido da "language understanding", ou seja de "reading machines" para "talking machines". As vantagens são destacadas pelos autores: "Annotated input of a synthesizer would allow a finer specification of speaking style and of the intended interpretation of a message. Such input is a prerequisite for speech-to-speech applications where not only linguistic but paralinguistic information should be translated and rendered properly. Proposals have already been put forward for menu-driven interfaces that allow for the switching of speaker, language, emotion, and speaking-style with automatic or semi-automatic adjustments to the mark-up of the input." (Bailly *et al.* 2003).

Além destes cinco tópicos, gostaríamos de acrescentar os seguintes:

1. **Síntese por HMMs** (Hidden Markov Models), primeiramente proposta por Tokuda *et al.* (1995) e que constitui o novo paradigma nas técnicas de síntese (Tokuda, 2004); trata-se da aplicação da mais popular técnica usada para o Reconhecimento de voz na Síntese da Fala. A descrição e demonstrações do sistema podem ser encontradas na página oficial do HTS[70] e no capítulo de Taylor (2007) sobre "HMM synthesis".

2. **"Voice conversion",** também designado "voice transformation" e "voice morphing", que consiste na transformação da voz de um determinado locutor (*source speaker*) na voz de outro locutor distinto (*target speaker*), através da manipulação das suas características. São muitos já os trabalhos relacionados (só para citar alguns exemplos: Turk *et al.* 2005; Uto *et al.* 2006; Percybrooks & Moore II, 2007; Erro & Moreno 2007).

---

[67] Dissertação de doutoramento intitulada: "*Talking heads: Models and applications for multimodal speech synthesis*"

[68] Um dos oradores convidados do Interspeech 2007.

[69] Vídeos disponíveis em: http://www.speech.kth.se/multimodal/video/index.html (08-12-2007).

[70] Disponível em: http://hts.sp.nitech.ac.jp/ (08-12-2007).

3. **"Speaker adaptation"** que tem por objectivo chegar a um sinal de fala diferente do que o sintetizador-base produziria. Porém, e ao contrário da voice conversion, em que se altera o sinal de fala após a síntese, neste caso é o próprio engine que é alterado e adaptado para um novo falante (Zen *et al.,* 2006). Veja-se ainda as demontrações disponíveis na página de K. Tokuda[71].

4. **Novas aplicações** usando sistemas de TTS, com destaque para as destinadas à promoção da acessibilidade (Nakamura, 2007) e para a adaptação a ambientes móveis (veja-se a título de exemplo o *Voice command* do *Windows Mobile*[72]).

Existem ainda muitos problemas não resolvidos ao nível da conversão grafema-fone(ma) e do processamento de texto, o que justifica a permanência desse tópico ainda no Interspeech 2007. Além disso, no que diz respeito à conversão texto-fala de línguas com menos recursos linguísticos e económicos ou com mercados mais reduzidos, existe ainda um longo percurso a percorrer em todos os níveis envolvidos.

Até este momento, temos apresentado o panorama internacional da área da Síntese da Fala. Na geografia da lusofonia, por seu lado, a Síntese da Fala possui pouco mais de uma década. Historicamente, podemos situar o início da Síntese da Fala do português em 1991 (Oliveira, *et al.*, 1991), data da publicação sobre o primeiro TTS para PE, o DIXI, apesar de outros grupos estarem já a despontar também para a área. Apesar da juventude da disciplina, têm sido feitos grandes esforços de acompanhamento do estado da arte internacional, o que tem resultado numa comunidade científica muito activa e em afirmação no plano internacional.

Apesar do interesse crescente pelas tecnologias de fala por parte da comunidade científica lusófona, ainda são muito poucos os grupos que trabalham em Síntese da Fala do Português, uma vez que a grande maioria prefere dedicar-se ao Reconhecimento de Voz. Além disso, trata-se de uma comunidade ainda muito dependente do financiamento estatal, o que em parte justifica a sua pequena dimensão. Sendo assim, destacamos os seguintes grupos dedicados à Síntese da Fala:

- Em Portugal:
    - L2F – Spoken Language Systems Lab[73], INESC-ID/Insituto Superior Técnico, Lisboa, reunindo um grupo interdisciplinar com mais de uma década de experiência na área;

[71] http://www.sp.nitech.ac.jp/~tokuda/HTS_demo/speaker_adaptation/index.html (08-12-2007).
[72] Trata-se de uma funcionalidade do Windows Mobile que permite interagir consultar a agenda, os contactos e fazer chamadas, entre outras coisas. Mais informações em: http://www.microsoft.com/windowsmobile/voicecommand/default.mspx (08-12-2007).
[73] Link para o website: https://www.l2f.inesc-id.pt/wiki/index.php/Main_Page (08-12-2007).

- LSS – Laboratory of Signals and Systems[74], Faculdade de Engenharia da Universidade do Porto, um dos primeiros grupos depois do L2F a desenvolver sintetizadores de fala em PE;
- Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA)[75], em cooperação com o Departamento de Línguas e Culturas[76], Universidade de Aveiro;
- MLDC – Microsoft Language Development Center[77], um centro de investigação e desenvolvimento da Microsoft, em Porto Salvo, onde se desenvolvem presentemente sistemas de conversão texto-fala para PE e PB;

- No Brasil:
  - LPS – Laboratório de Processamento de Sinais, da Universidade Federal do Rio de Janeiro[78];
  - LINSE – Laboratório de Circuitos e Processamento de Sinais, da Universidade Federal de Santa Catarina[79];
  - Speech Prosody Studies Group[80], que reúne investigadores da Universidade Estadual de Campinas, da Pontifícia Universidade Católica de São Paulo e da Universidade Federal de Minas Gerais;
  - Laboratório de Processamento Digital da Fala - LPDF[81], Departamento de Comunicações (DECOM) da Faculdade de Engenharia Elétrica e de Computação (FEEC), em colaboração com o LAFAPE - Laboratório de Fonética e Psicolingüística do Instituto de Estudos da Linguagem (IEL), ambos os grupos pertencentes à Universidade Estadual de Campinas (UNICAMP), um dos grupos mais antigos do Brasil.

[74] Link para o website: http://www.fct.mctes.pt/unidades/index.asp?p=3&u=720 (08-12-2007).
[75] Link para o website: http://www.ieeta.pt/ (08-12-2007).
[76] Link para o website: http://www2.ii.ua.pt/cidlc/gcl/index.htm (08-12-2007).
[77] Link para o website: http://www.microsoft.com/portugal/mldc/default.mspx (08-12-2007).
[78] Link para o website: http://www.lps.usp.br/ (08-12-2007).
[79] Link para o website: http://www.linse.ufsc.br/index.php?language=pt-BR (08-12-2007).
[80] Link para website: http://www.experimentalprosodybrazil.org/ (08-12-2007).
[81] Website disponível em: http://www.decom.fee.unicamp.br/lpdf/ (08-12-2007).

A nível académico, estão descritos vários sintetizadores de fala em PE, como os sintetizadores por formantes DIXI (Oliveira, 1996), continuado no projecto DIXI+[82] e presentemente melhorado no âmbito do projecto Tecnovoz, actualmente dispondo de duas vozes masculinas e duas femininas[83], e o Multivox (Teixeira, 1995; Teixeira & Freitas 1998); o sintetizador de base articulatória desenvolvido pela Universidade de Aveiro (Teixeira, 2000), sintetizadores por concatenação de unidades (Barros, 2001; Carvalho *et al.*, 2003) ou o sintetizador usando a síntese por HMMs (Barros *et al.*, 2005).

Para o PB, são de salientar o *Aiuruetê*, o sintetizador concatenativo do LPDF-DECOM da UNICAMP, iniciado em 1991, que tem sido alvo de sucessivos desenvolvimentos, (Boeffard & Violaro, 1994; Violaro & Böeffard, 1998; Gomes, 1998; Barbosa *et al.*, 1999; Simões *et al.*, 2000), o sintetizador concatenativo desenvolvido pelo LINSE (Nicodem *et al.*, 2005; Nicodem *et al.*, 2007), e o sintetizador baseado em HMMs[84] desenvolvido pelo LPS em colaboração com o Nagoya Institute of Technology (Maia *et al.*, 2006; Maia, 2006).

Relativamente à comunidade que se dedica à Síntese da Fala em galego, é possível identificar dois grupos que trabalham em estreita colaboração: o Centro Ramón Piñeiro para a Investigación en Humanidades[85], na área da Linguística, e o Grupo de Tecnoloxías do Sinal da Universidade de Vigo[86], na área da Engenharia. Desta actividade conjunta resultou um sintetizador concatenativo para galego, o Cotovía[87]. Para mais detalhes sobre o sistema veja-se Banga *et al.* (2001), González (2004), Campillo *et al.* (2005), Banga *et al.* (2006). O galego dipõe também de um

---

[82] Projecto desenvolvido pelo L2F em cooperação com o CLUL – Centro de Linguística da Universidade de Lisboa: http://www.speech.inesc.pt/~lco/dixiplus/ (08-12-2007).

[83] O projecto Tecnovoz é o maior projecto nacional alguma vez realizado neste domínio tecnológico, actualmente activo, que reúne vários grupos académicos (INESC- ID, INESC Inovação, Universidades de Coimbra e do Minho) e empresariais (Companhia Portuguesa de Computadores – Healthcare Solutions, Anditec - Tecnologias de Reabilitação, Lda., Datelka - Engenharia e Sistemas, Lda., EDISOFT – Empresa de Serviços e Desenvolvimento de Software, Priberam Informática, Promosoft SIS – Software de Sistemas, Rádio e Televisão de Portugal e Tecmic - Tecnologias de Microelectrónica), com um objectivo de desenvolver produtos de competitividade comercial integrando sistemas de diálogo e tecnologias de voz, dos quais destacamos alguns exemplos: sistemas de monitorização telefónica, arquivos de voz e relatórios, desktop médico, soluções de corretagem, bancárias e de seguros, legendagem e transcrição automáticas, sistema de comunicação aumentativo, etc. (http://www.tecnovoz.pt/web/?id=4&mid=4, 15-02-2008. Mais informações sobre o Tecnovoz disponíveis em: http://www.tecnovoz.com.pt/web/?id=756&mid=1 (08-12-2007). Informação sobre o DIXI no âmbito do projecto Tecnovoz disponível em: https://tecnovoz.l2f.inesc-id.pt/oquefazemos/motores/index.php e demonstração disponível em: https://tecnovoz.l2f.inesc-id.pt/demos/tts/index.php (15-02-2008).

[84] Teste online disponível em http://kt-lab.ics.nitech.ac.jp/~maia/demo.html (08-12-2007)..

[85] Descrição do projecto disponível em: http://www.cirp.es/prx2/sinte.html (08-12-2007).

[86] Link para website: http://www.gts.tsc.uvigo.es/web/index.php?cGF4aW5hPSZwYXJhbTE9MjU= (08-12-2007).

[87] Demonstração do sistema disponível em: http://www.gts.tsc.uvigo.es/cotovia/cotovia.gl.html (08-12-2007).

sistema comercial vendido pela empresa catalão Verbio Speech Technologies com uma voz masculina[88].

Os tópicos que actualmente suscitam maior interesse por parte da comunidade científica prendem-se com a conversão grafema-fone e com a silabificação (Oliveira *et al.*, 2004; Oliveira *et al.*, 2005; Teixeira *et al.*, 2006; Barbosa *et al.*, 2003a), com a prosódia e suas interfaces com a fonologia, sintaxe ou pragmática (Barbosa, 2006; Teixeira, 2004; Braga & Marques, 2004; Seara *et al.,* 2007), com avaliação (Weiss *et al.*, 2007), com a síntese de emoções (Cabral, 2006; Cabral & Oliveira, 2006), com a "voice conversion" (Weiss, 2007), com a síntese por HMMs (Maia, 2006; Maia *et al.*, 2007) e com a síntese multi-modal (Martino, 2005; Raimundo *et al.*, 2007; Martino & Violaro, 2007).

Apesar desta variedade de sistemas, poucos autores revelam os algoritmos de processamento da linguagem natural envolvidos no seu TTS. Nesta dissertação, descreveremos esses algoritmos, apresentaremos o teste do seu desempenho com textos extraídos de corpora autênticos e discutiremos as suas aplicações e limitações.

Finalmente, o português começa a ser uma língua que suscita interesse comercial, dada a dimensão do seu mercado (cerca de 235 milhões de pessoas), o que explica que empresas como a Nuance, a Loquendo e a Acapela comercializem TTSs nas duas variedades do português (PE e PB), que a Microsoft também os esteja a desenvolver para PE e PB e que outras empresas apostem no desenvolvimento do português de acordo com a proximidade do seu mercado, ou seja, a americana Aculab dispõe de um TTS em português do Brasil, enquanto que a catalã Verbio Speech Technologies aposta no português europeu.

1.3. Arquitectura do sistema

Apesar da grande flutuação existente a nível da arquitectura dos modernos sistemas de síntese da fala[89], existem pelo menos três blocos comuns a todos eles: o

[88] Disponível para teste em: http://www.verbio.com/webverbio2/html/demos_ttsonline.php (08-12-2007).

[89] Taylor (2008: 38-41) apresenta uma tipologia de 8 modelos de TTSs baseada nas diferentes arquitecturas possíveis. Apesar de mencionar que a maioria dos sistemas combina mais do que um modelo, apresenta as seguintes classes de sistemas de conversão Texto-Fala: *1) common form model* ("In the common form model, there are essentially two components; a text analysis system which decodes the text signal and uncovers the form, and a speech synthesis system which encodes this form as speech."); *2) signal-to-signal model* ("In this model, the process is seen as one of converting the written signal into a spoken one directly."); *3) pipelined models* ("Each module performs one specific task such as part-of-speech tagging, or pause insertion and so on. No explicit distinction is made between analysis and synthesis tasks… Often the modules are not explicitly linked so that different theories and techniques can co-exist in the same overall system."); *4) text as language models* ("The text itself is taken as the linguistic message, and synthesis is performed from this. As the text is rarely clean or unambiguous enough for this to happen directly, a text

pré-processamento de texto ou *front-end*, o motor de síntese ou *back-end* e a base de dados de voz ou *voice font*. Em Huang *et al.* (2001, Parte IV, capítulo 14) apresenta-se detalhadamente os vários módulos subjacentes a um sistema de conversão texto-fala e discute-se os principais problemas e desafios envolvidos em cada módulo. Em todos os casos, o objectivo é a geração de fala sintética resultante da conversão do texto em etiquetas fonéticas, obtida à saída do *front-end*, etiquetas essas que serão depois descodificadas e transformadas pelo motor de síntese em voz.

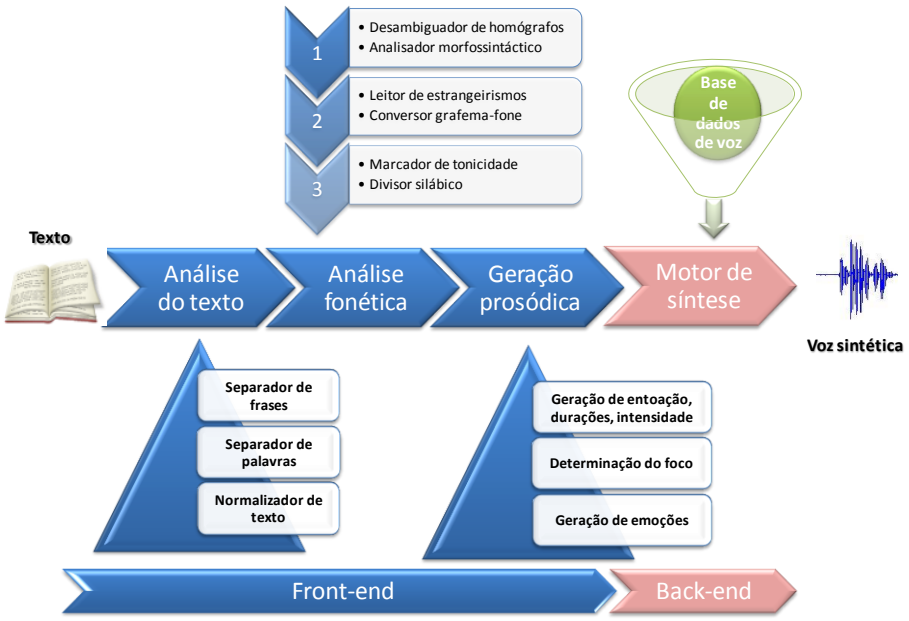

Figura 1: Arquitectura do nosso TTS.

---

normalization process is normally added, as a sort of pre-processor to the synthesis process itself."); *5) grapheme and phoneme form models* ("This approach is in many ways similar to the common form model in that first a grapheme form of the text input is found, and this in then converted to a phoneme form for synthesis.") – modelo no qual se enquadra o nosso sistema; *6) full linguistic analysis models* ("Some systems go much further in terms of linguistic analysis and perform morphological analysis, part-of-speech tagging and syntactic parsing."); *7) complete prosody generation* ("the F0, phrasing, stress and so on in an utterance are all directly determined by prosody, and so to generate speech we have to generate all these quantities with an explicit prosodic model."); *8) prosody from the text* ("A common assumption is that the text does in fact contain enough information to determine prosody, and so many TTS systems have modules which try and predict prosodic representations directly from the text, often with the assumption that this is an analysis process with a right and wrong answer.").

Na Figura 1, apresenta-se a arquitectura do nosso sistema. O *front-end* constitui um dos módulos mais complexos e é composto por três componentes:

1. a análise de texto, do qual fazem parte o separador de frases (responsável por identificar o tipo de frase, informação que pode ser útil para a geração dos modelos prosódicos), o separador de palavras e o normalizador de texto (em que se incluem os sub-módulos de interpretação de pontuação, expansão de abreviaturas, leitura de siglas e de acrónimos e conversão de numerais, datas, números romanos e árabes, quantias em dinheiro, números ordinais);

2. a análise fonética, da qual fazem parte os módulos de leitura de estrangeirismos, de desambiguação de homógrafos heterófonos, de análise morfossintáctica e de conversão grafema-fone,

3. análise e geração prosódica, módulo que aproveita as informações anteriormente obtidas a partir do texto, como tipo de frase, divisão silábica, marcação de tónica, classificação morfossintáctica, para a geração de contornos prosódicos.

O *back-end* é composto pelo motor de síntese que interpreta a transcrição fonética gerada pelo *front-end* e a transforma em fala sintética. Neste trabalho, estamos a usar a técnica de síntese por HMMs, alimentada por uma base de dados de 1000 frases foneticamente balanceada, com a duração de cerca de 1 hora e 15 minutos de voz (com silêncios). Esta técnica permite precisamente obter bons resultados com uma base de dados relativamente reduzida. A base de dados de voz foi rigorosamente seleccionada, gravada em estúdio profissional e etiquetada automaticamente com verificação manual. Este assunto será desenvolvido no Capítulo 6.

Na presente dissertação, propõem-se novos módulos de análise de texto e de análise fonética através de uma metodologia de análise linguística. Estes módulos representam a base para qualquer sistema de conversão de Texto-Fala em português independentemente da técnica de síntese que venha a ser usada. Estes módulos fornecem ainda a base para o desenvolvimento de modelos de predição prosódica que usem a informação de tipo de frase, acento de palavra e sílaba.

Todo o sistema apresentado foi programado em linguagem C/C++ para ambiente Windows. A interface gráfica foi desenvolvida em Borland Delphi.

1.4. Síntese do capítulo 1

Neste capítulo, foram apresentados os fundamentos teóricos, estado da arte e arquitectura do sistema no qual se enquadra este trabalho. Em jeito de síntese, podemos concluir:

- A presente dissertação inscreve-se na área interdisciplinar da Linguística Aplicada à Síntese da Fala ou na área do Processamento da Linguagem Natural aplicado à Síntese da Fala, reunindo assim os fundamentos teóricos da Linguística e as metodologias de trabalho e validação dos resultados da Informática e da Engenharia;
- O estado da arte da Síntese da Fala é dominado actualmente pelas técnicas por concatenação e mais recentemente pela síntese por HMMs; são já inúmeros os sistemas comerciais disponíveis para várias línguas;
- O estado da arte da Síntese da Fala tem sofrido avanços muito rápidos e tem sido feito sobretudo a nível académico, podendo ser acompanhado através das publicações efectuadas em congressos internacionais de periodicidade anual como o Interspeech e o ICASSP, em workshops de Speech Synthesis promovidas pela ISCA ou em revistas da especialidade como *Computer Speech and Language*, *IEEE Transactions on Audio, Speech, and Language Processing* e *The ACM Transactions;*
- As temas que suscitam maior interesse no actual estado da arte da Síntese da Fala são: a avaliação, a síntese multi-língua, a síntese por emoções/expressiva, a síntese multi-modal ou síntese audio-visual, os inputs e sistemas de mark-up, a síntese por HMMs, a *Speaker adaptation* e as novas aplicações;
- No contexto lusófico, ainda são poucos os grupos dedicados à Síntese da Fala do Português; destaque para o esforço levado a cabo pelo consórcio português do Projecto Tecnovoz, actualmente em curso;
- Os tópicos que suscitam maior interesse por parte da comunidade científica são a conversão grafema-fone e a silabificação, a prosódia e suas interfaces com a fonologia, sintaxe ou pragmática, a avaliação a síntese de emoções, a "voice conversion", a síntese por HMMs e a síntese multi-modal;
- Apesar de alguma variedade de sistemas TTS descritos para o Português, poucos autores revelam os algoritmos de processamento da linguagem natural envolvidos no *front-end* do sistema;
- Em relação à arquitectura, este trabalho concentra-se no desenvolvimento de módulos de PLN destinados ao funcionamento mais eficaz do front-end de um TTS através de uma metodologia de análise linguística. Todo o sistema apresentado foi programado em linguagem C/C++ para ambiente Windows. A interface gráfica foi desenvolvida em Borland Delphi.

Capítulo 2

Pré-processamento de texto

O pré-processamento ou normalização de texto significa a conversão de toda a espécie de símbolos, abreviaturas, siglas, acrónimos, numerais, fórmulas matemáticas, etc. em sequências ortográficas adequadas para uma subsequente transcrição fonética. É uma das tarefas da Análise de Texto mais estudadas. É comum que este assunto seja resolvido por dicionários ou léxicos, ou seja, através de uma lista de entradas e da respectiva expansão ortográfica ou imediata transcrição fonética. Este módulo é dependente da língua, o que faz com que seja muito específico e que necessite de permanente actualização, dado que o léxico das línguas está em permanente expansão. Contudo, não é comum encontrar-se na literatura detalhes, tabelas, léxicos ou algoritmos sobre este módulo, sobretudo no que se refere ao pré-processamento do português. Neste capítulo, apresentaremos os dicionários e listas implementados e os algoritmos de conversão propostos para resolver vários problemas inerentes à tarefa de normalização do texto. Este módulo foi implementado e testado com corpora reais, tendo-se obtido 99,88% de acerto para o conversor de siglas/acrónimos e números romanos e 99,86% de acerto para o conversor de dígitos (numerais árabes cardinais e ordinais, datas, horas, números com casas decimais, medidas e pontuação desportiva). Discutiremos ainda a aplicabilidade deste módulo ao português do Brasil e ao galego.

2.1. Separador de frases

O pré-processamento começa com a separação do texto em frases. Esta separação é feita sempre que o sistema identifique sinais de pontuação forte, como ponto <.>, reticências <…>, ponto de interrogação <?> e ponto de exclamação <!>. Seguidamente, foram identificadas as palavras seguidas de ponto que não representam final de frase, como é o caso de algumas abreviaturas (ex. <a.C.> → <antes de Cristo>, <Av.> → <Avenida>) e de formas de tratamento ou títulos (ex. <Eng.º> → <Engenheiro>, <Sr.> → <Senhor>). Neste ponto, o sistema apenas ignora a pontuação destas palavras e não termina a frase no sinal de pontuação que as termina. No total, foram identificadas 334 abreviaturas com ponto recolhidas a partir de Estrela *et al.* (2004) e Bergström & Reis (2007). De igual forma, não são considerados separadores de frase letras do alfabeto, números árabes ou números romanos seguidos

de ponto, quando usados como forma de listagem de itens. Mais adiante, neste capítulo, trataremos da expansão de abreviaturas.

2.2. Separador de palavras

Este módulo tem apenas como função separar as palavras dentro das frases do texto que servir de input. À partida, trata-se de uma tarefa simples, já que as palavras em português aparecem separadas por espaços em branco. Apenas há que ter em consideração os casos de palavras separadas por hífen, como as palavras compostas (ex. <cor-de-rosa>, <guarda-chuva>), as formações com prefixos (ex. <antí-míssil>, <ex-marido>), em formações com preposições (ex. <há-de>) ou as formas verbais com clíticos em mesóclise (ex. <di-lo-ás>, <dir-me-ias>) ou ênclise (ex. <diz-se>, <pago-lho>), que devem ser tratados como palavras separadas e autónomas. Mesmo os clíticos, que são palavras átonas, e por isso se subordinam à estrutura fonológica das formas verbais, são tratados como palavras independentes, embora listados de forma a serem reconhecidos no momento da conversão grafema-fone.

2.3. Conversor de símbolos e caracteres especiais

Ao nível da anotação fonética, seguimos ao longo da presente dissertação o alfabeto SAMPA[90] (vide Tabela 1) com uma extensão (a consoante lateral velarizada [l*] que ocorre na articulação da palavra <sal> em PE), por ser o alfabeto mais adequado e amplamente usado no processamento computacional das línguas.

Considerámos símbolos todos os caracteres simples (que ocupam apenas um espaço) que não sejam números árabes nem romanos, que não sejam sinais de pontuação e que não pertençam ao alfabeto português, nem que sejam letras estrangeiras que podem ser encontradas em estrangeirismos portugueses, nomeadamente <k>, <w> e <y>. São símbolos os caracteres especiais que constam dos teclados, sinais relativos às operações aritméticas, letras gregas etc., como se pode ver na Tabela 2, em que estão representados os mais usuais. A sílaba tónica é assinalada com o algarismo <1> imediatamente depois da vogal tónica.

[90] Acerca do SAMPA: "SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-89 by an international group of phoneticians, and was applied in the first instance to the European Communities languages Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese, and Spanish (1993)." Além disso, "Where Unicode (ISO 10646) is not available or not appropriate, SAMPA and the proposed X-SAMPA (Extended SAMPA) constitute the best robust international collaborative basis for a standard machine-readable encoding of phonetic notation." Disponível em http://www.phon.ucl.ac.uk/home/sampa/index.htm.

O conversor de símbolos e caracteres especiais é percorrido sempre que um texto entra no sistema, procurando assim encontrar correspondências. Se o símbolo se encontrar na lista o sistema devolve a transcrição fonética imediata. O conversor de símbolos possui uma lista aberta, que pode ser expandida a todo o momento.

Tabela 1: Alfabeto SAMPA para o português.

Símbolos SAMPA	Tipologia
[p], [t], [k]	**oclusivas orais surdas**
[b], [d], [g]	**oclusivas orais sonoras**
[m], [n], [J]	**oclusivas nasais**
[f], [s], [S]	**fricativas surdas**
[v], [z], [Z]	**fricativas sonoras**
[l], [L], [l*]	**laterais**
[r], [R]	**vibrantes**
[a], [6], [E], [e], [@], [O], [o], [i], [u]	**vogais orais**
[6~], [e~], [o~], [i~], [u~]	**vogais nasais**
[j], [w], [j~], [w~]	**semivogais**

Tabela 2: Símbolos, sua designação e transcrição fonética.

Símbolo	Conversão ortográfica	Transcrição Fonética
#	cardinal	k6r.di.na1l*
$	dólar	dO1.lar
%	por cento	pur.se~1.tu
&	e comercial	i.ku.m@r.sja1l*
*	asterisco	6S.t@.ri1S.ku
+	mais	ma1jS
-[91]	menos	me1.nuS
/[92]	barra à direita, ou	ba1.Ra.di.r61j.t6, o1w
=	igual a	i.gwa1.l6
@	arroba	6.Ro1.b6
\	barra à esquerda	ba1.Ra.@S.ker1.d6
_	underscore	6~.d6r.skO1.r@
~[93]	til	ti1l*
£	libra	li1.br6
¥	yen	jE1n
€	euro	e1w.ru
©	copyright	kO.pi.Ra1jt

[91] Este sinal matemático pode ser confundido com hífen. Lê-se <menos> quando está entre dígitos. Caso contrário não se lê.

[92] O output depende do contexto. Lê-se "barra à direita" se estiver em urls, começados por "http" ou "www". Lê-se "ou" quando separa duas palavras (ex. m/f – masculino ou feminino).

[93] O output é definido consoante o contexto. O til só é lido quando se encontra sozinho ocupando um espaço. É comum em links para páginas web ou urls. Não é lido quando aparece sobre uma vogal, sendo aqui interpretado como marca de nasalidade.

Tabela 2: Símbolos, sua designação e transcrição fonética (continuação).

®	marca registada	ma1r.k6.R@.ZiS.ta1.d6
° C[94]	grau(s) Celsius	gra1w/S. sE1l*.siwS
÷	a dividir por	6.di.vi.di1r.pur
×	vezes	ve1.z@S
≤	menor ou igual a	m@.nO1r.ow.i.gwa1.l6
≥	maior ou igual a	maj.O1r.ow.i.gwa1.l6
≠	diferente de	di.f@.re~1.t@
∞	infinito	i~.fi.ni1.tu
μ	miu	mju1
α	alpha	a1l*.f6
β	beta	bE1.t6
Γ, γ	gamma	g61.m6
Δ, δ	delta	dE1l*.t6
ε	epsilon	E1psi.lOn
η	eta	E1.t6
ζ	zeta	zE1.t6
θ	teta	tE1.t6
ι	iota	jO1.t6
Λ,λ	lambda	l6~1.bd6
ν	niu	nju1
ξ	xi	Si1
Π, π	pi	pi1
ρ	ro	RO1
Σ, σ	sigma	si1.gm6
τ	tau	ta1w
φ	phi	fi1
χ	chi	Si1
ψ	psi	psi1
ω	ómega	O1.m@.g6
3[95]	cúbico(s)	ku1.bi.ku(S)
2[96]	quadrado(s)	kw6.dra1.du(S)

2.4. Expansor de abreviaturas

Considerámos abreviaturas as palavras truncadas terminadas por ponto (ex. <adj.>, <cit.>), ou não terminadas por ponto (como as abreviaturas bíblicas ou os símbolos químicos) os grafemas simples (ex. como os pontos cardeais <N>, <S>, <E>, <O>) ou sequências de grafemas oscilando entre maiúsculas e minúsculas finalizados por um ponto ou barra (ex. <a.C>, <A/C>) e/ou terminados por caracteres em expoente

[94] Este símbolo tem uma regra associada: "O carácter anterior tem que ser um dígito. Se o carácter anterior for um dígito =1 então a transcrição fonética =[graw], se for um dígito ≠ 1 então lê-se [grawS]. Decidimos usar a unidade Celsius por ser mais comum nos textos portugueses de tipo jornalístico, embora haja outras unidades de medida de temperatura: Fahrenheit, Kelvin.

[95] No caso dos expoentes, aplica-se a regra: "Lê-se [kubikuS] quando os caracteres anteriores forem uma das seguintes sequências: <ml, cm, dm, m, km, ml, cl, dl, l, kl e quando o carácter anterior à unidade for um dígito ≠ 1"; se for =1, então lê-se [kubiku].

[96] Este caso é semelhante anterior, apenas o que muda é a saída, que neste caso é [kw6dradu/S].

(como em títulos académicos: <eng.º>, <dr.ª>). A Tabela 3 apresenta as abreviaturas mais comuns (Bergström e Reis, 2007:96-97) que foram consideradas no nosso módulo de pré-processamento.

Tabela 3: Abreviaturas, sua expansão e transcrição fonética.

Abreviatura	Expansão	Transcrição Fonética
a.C.	antes de Cristo	6~1.t@Z.d@.ˈkri1S.tu
AA.VV	autores	aw.to1.r@S
A/C, a/c	ao cuidado	aw.kuj.da1.du
abrev.	abreviatura	6.br@.vi.6.tu1.r6
adj.	adjectivo	6.djE.ti1.vu
adv.	advérbio	6.dvE1r.bju
art.º	artigo	6r.ti1.gu
Av.	avenida	6.v@.ni1.d6
cf.	confira	ko~.fi1.r6
cit.	citação	si.t6.s6~1w~
cl	centilitro	se~.ti.li1.tru
cm	centímetro	se~.ti1.m@.tru
conj.	conjunção	ko~.ju~.s6~1w~
cv.	cave	ka1.v@
D.[97]	dom/dona	do~1, do1.n6
d.C.	depois de Cristo	d@.po1jZ.d@.kri1S.tu
d.to	direito	di.r61j.tu
dep.	departamento	d@.p6r.t6.me~1.tu
dl	decilitro	d@.si.li1.tru
dm	decímetro	d@.si1.m@.tru
dr.	doutor	do.to1r
dr.ª	doutora	do.to1.r6
E	este	E1S.t@
e.C.	era cristã	E1.r6.kriS.t6~1
ed.	edifício	i.di.fi1.sju
eng.º	engenheiro	e~.Z@.J61j.ru
eng.ª	engenheira	e~.Z@.J61j.r6
esq.	esquerdo	@S.ke1r.du
etc.	et cetera	E.t@.sE1.t@.r6
Ex.ma	Excelentíssima	6jS.s@.le~.ti1.si.m6
Ex.mo	Excelentíssimo	6jS.s@.le~.ti1.si.mu
fasc.	fascículo	f6S.si1.ku.lu
fl.	folha	fo1.L6
i.e.	isto é	i1S.tu.E
km	quilómetro	ki.lO1.m@.tru
km2,km^2	quilómetro quadrado	ki.lO1.m@.tru.kwa.dra1.du
Lda.	limitada	li.mi.ta1.d6
m	metro	mE1.tru
m2,m^2	metro quadrado	mE1.tru.kwa.dra1.du
M.ª	Maria	m6.ri1.6
M.el	Manuel	m6.nu.E1l*
m.q.	o mesmo que	u.me1Z.mu.k@
maj.	major	m6.ZO1r

[97] Regra para Dom: Se P+1= termina por <–o> → [do~]. Regra para Dona: Se P+1= termina por <–a> → [don6]. Para tratar nomes próprios terminados por consoante (ex. <Leonor>) ou por <-e> (ex. <Filipe>), é necessário um analisador morfológico que anote a informação de género. Trata-se de um assunto a ser contemplado em trabalho futuro.

Tabela 3: Abreviaturas, sua expansão e transcrição fonética (continuação).

mar.	marechal	m6.r@.Sa1l*
mg	miligrama	mi.li.gr61.m6
ml	mililitro	mi.li.li1.tru
mm	milímetro	mi.li1.m@.tru
N	norte	nO1r.t@
nº	número	nu1.m@.ru
N.B.	note bem	nO1.t@.b6~1j~
NE	nordeste	nOr.dE1S.t@
NO	noroeste	nO.ru.E1S.t@
N.S.	Nosso Senhor	nO1.su.s@.Jo1r
O	oeste	O.´E1S.t@
obs.	observação	Ob.s@r.v6.s6~1w~
P.e	padre	pa1.dr@
P.M.P.	por mão própria	pur.m6~1w~.prO1.prja
P.S.	*post scriptum*	pO1st.skri1.ptu~
pág.	página	pa1.Zi.n6
p.	página	pa1.Zi.n6
pç.	praça	pra1.s6
pct.	praceta	pr6.se1.t6
pl.	plural	plu.ra1l*
prep.	preposição	pr@.pu.zi.s6~1w~
prof.	professor	pru.f@.so1r
pron.	pronome	pru.no1.m@
q.b.	quanto baste	kw~6~1.tu.ba1S.t@
R.S.F.F	responda se faz favor	R@S.po~1.d6.s@.fa1S.f6.vo1r
r/c	rés-do-chão	RE1Z.du.S6~1w~
ref.	referência	R@.f@.re~1.sj6
rep.	repartição	R@.p6r.ti.s6~1w~
rev.	reverendo	R@.v@.re~1.du
S.	são	s6~1w~
S.to	santo	s6~1.tu
S	sul	su1l*
s.	segundo	s@.gu~1.du
SE	sueste	su.E1S.t@
sing.	singular	si~.gu.la1r
SO	sudoeste	su.du.E1S.t@
Sr.	senhor	s@.Jo1r
Sr.a	senhora	s@.Jo1.r6
S.S.	Sua Santidade	sw6.s6~.ti.da1.d@
SS.	Santíssimo	s6~.ti1.si.mu
subst.	substantivo	subS.t6~.ti1.vu
tít.	título	ti1.tu.lu
tr.	transitivo	tr6~.zi.ti1.vu
urbaniz.	urbanização	ur.b6.ni.z6.s6~1w~
v.	verbo	vE1r.bu
V.A.	Vossa Alteza	vO1.sal*.te1.z6
v.g.	*verbi gratia*	vE1r.bi.gra1.tja
vd.	*vide*	vi1.dE
visc.	visconde	viS.ko~1.d@
v.s.f.f.	vire se faz favor	vi1.r@. s@.fa1S.f6.vo1r
V.M.	Vossa Majestade	vO1.s6.m6.Z@S.ta1.d@

Foi também reunida uma lista de abreviaturas sem ponto, constuída por 73 abreviaturas bíblicas (ex. <Ex - Êxodo>, <Sl – Salmos>) e 104 abreviaturas químicas (ex. <ag- prata>, <al - alumínio>), que não incluímos aqui por razões de espaço. Além

disso, algumas destas abreviaturas necessitam de desambiguação por contexto, o que não foi ainda considerado neste trabalho (ex. <Ag – Profeta Ageu> em sentido bíblico e <Ag - prata> em sentido químico) por limitações temporais. O expansor de abreviaturas possui uma biblioteca com um léxico e sua respectiva transcrição fonética. Quando recebe o texto como input, este sub-módulo de pré-processamento, à semelhança do que acontece com o conversor de símbolos e caracteres especiais, faz uma verificação da existência de abreviaturas no léxico de pré-processamento e se encontra uma correspondência produz a sua transcrição imediata.

2.5. Leitor de siglas e acrónimos

Apesar de alguma flutuação existente na designação de sigla na literatura da especialidade[98], distinguimos *Sigla* de *Acrónimo,* na medida em que no primeiro caso a sequência fonológica não permite formar sílabas válidas (ex: <ACP>, <PSD>), apenas permitindo soletrar a palavra, enquanto no segundo caso, a sequência fonológica permite formar sílabas válidas, o que permite "ler" a palavra (ex. <APEL>, <ANJE>).

Mendes *et al.* (2004), baseados em estudos anteriores (Viana *et al.,* 1994 e Trancoso & Viana, 1997), num trabalho enquadrado no ensino do português como

[98] A definição de sigla e de acrónimo não parece estar clara na literatura, embora a acronímia seja em si mesma um processo morfológico de formação de novas palavras. Veja-se por exemplo as definições do CD-ROM intitulado *Terminologia Linguística para os Ensinos Básico e Secundário*, da responsabilidade do Ministério da Educação (Março de 2005): **Sigla:** «1. Processo morfológico consistindo na redução de uma palavra ou de um grupo de palavras às suas iniciais para designar organismos, partidos políticos, associações, clubes desportivos, etc. 2. Letra inicial ou grupo de letras iniciais que entram na composição da abreviação de certas palavras. Exemplos: SMAS – Serviços Municipalizados de Água e Saneamento; APET – Associação Portuguesa de Empresas de Tradução; PSD – Partido Social-Democrata; SCP – Sporting Clube de Portugal.» **Acronímia**: «Conjunto de processos morfológicos que levam à formação de acrónimos, cuja particularidade é a de serem pronunciados como uma palavra corrente; apresentam-se sob a forma de siglas, amálgamas ou de novas unidades lexicais. A acronímia é um fenómeno neológico muito produtivo em língua, com grande dinâmica, por exemplo, na atribuição de nomes próprios para designar novos objectos, instituições ou organismos. Exemplo: FENPROF – FEderação Nacional de PROFessores.». Repare-se ainda nas definições apresentadas pelo *Dicionário de Termos Linguísticos da Associação Portuguesa de Linguística e do Instituto de Linguística Teórica e Computacional* (1992, vol. II; pp. 16, 17 e 345), em que se diz que a pronúncia da sigla tanto pode ser soletrada como silábica: **Sigla** – «Termo complexo abreviado ou nome formado a partir das letras iniciais dos seus elementos. Uma sigla forma uma sequência cuja pronúncia é alfabética, silábica ou ambas. Exemplos: CEE, EDP.» **Acrónimo** – «Termo complexo abreviado, formado de letras ou grupos de letras de uma palavra ou sequência de palavras, que se pronuncia como uma palavra. Exemplos: EPAL, EUROTRA.» EPAL significa Empresa Pública das Águas Livres, EUROTRA significa Programa Europeu de Tradução Automática.»

Língua Estrangeira, sintetizam algumas regras importantes que permitem prever à pronúncia de siglas. A estrutura e a extensão são os parâmetros mais importantes para a leitura das siglas, que os autores designam por "acrónimos" no excerto que se segue:

> A extensão e a estrutura são factores preponderantes no que diz respeito ao processo de pronunciação dos acrónimos.
> 1. Relativamente à estrutura, observa-se que a existência exclusiva de vogais ou de consoantes impede a leitura dos acrónimos, independentemente da sua extensão. Assim, são sempre soletrados acrónimos do tipo UA (Universidade de Aveiro), AIEA (Agência Internacional de Energia Atómica) ou CP (Caminhos de Ferro Portugueses), CGTP (Confederação Geral dos Trabalhadores Portugueses). Para serem lidos, os acrónimos necessitam de compreender, pelo menos, um grupo CV que corresponde ao tipo silábico mais frequente do português.
> 2. No que diz respeito à extensão, são sempre soletrados os acrónimos constituídos por menos de três letras, como, por exemplo, RR (Rádio Renascença), BD (Banda Desenhada) ou PS (Partido Socialista). No mesmo sentido, e de acordo com Trancoso e Viana (1997), são lidos todos os acrónimos com mais de 5 caracteres, desde que possuam uma vogal. (Mendes *et al.*, 2004:14)

Trancoso e Viana (1995) atestam que o sistema de predição de siglas para o português, baseado em regras, apresentou 95% de acerto após a implementação de algumas regras mais contendo outros padrões fonológicos de leitura de siglas. São apresentados padrões mas não são apresentadas as regras que estão por trás deste trabalho.

Trancoso & Viana (1997), em outro trabalho, apresentam uma proposta para a leitura de siglas e acrónimos para 3 línguas, entre as quais o português, com resultados de 0,6% de erro para o português. Apresentam-se desta vez 9 regras para as 3 línguas, de carácter muito geral e sem apresentar as excepções referidas no artigo.

Em Barbosa *et al.* (2003b), apresenta-se uma proposta de resolução da questão da leitura de siglas e acrónimos[99] para PB, através de regras, com uma taxa de sucesso de 98,05%. Baseámo-nos nesta proposta para produzir o nosso algoritmo de leitura de siglas e acrónimos, apresentado na Figura 2.

Os candidatos do leitor de siglas e acrónimos são apenas as palavras constituídas na sua totalidade por letras maiúsculas, excepto letras e sequências de letras usadas na numeração romana (ver conversão de numerais). Embora haja acrónimos que também se encontrem grafados com letras maiúsculas e minúsculas (ex. <Cenjor – **Cen**tro de formação para **Jor**nalistas>, <Prodep – **Pro**grama de **D**esenvolvimento da **E**ducação em **P**ortugal>, <Fenprof – **Fe**deração **N**acional dos **Prof**essores>, Bergström e Reis, 2007:90-95), o nosso sistema interpreta-os como palavras, transferindo a sua conversão para o transcritor grafema-fone (descrito no Capítulo 6).

Na Figura 2, apresenta-se o funcionamento do nosso algoritmo de leitura de siglas/acrónimos. Primeiramente, foram definidas as seguintes condições: 1) são vogais: <a, e, i, o, u>; 2) são consoantes: <b,c,ç,d,f,g,h,j,k,l,m,n,p,q,r,s,t,v,w,x,y,z>.

---

[99] Note-se que em português do Brasil não se faz a distinção entre a designação de sigla e acrónimo, como em português europeu. A palavra acrónimo não é usada com o mesmo significado. Sigla designa simultaneamente FAB (Força Aérea Brasileira) e SBT (Sociedade Brasileira de Telecomunicações).

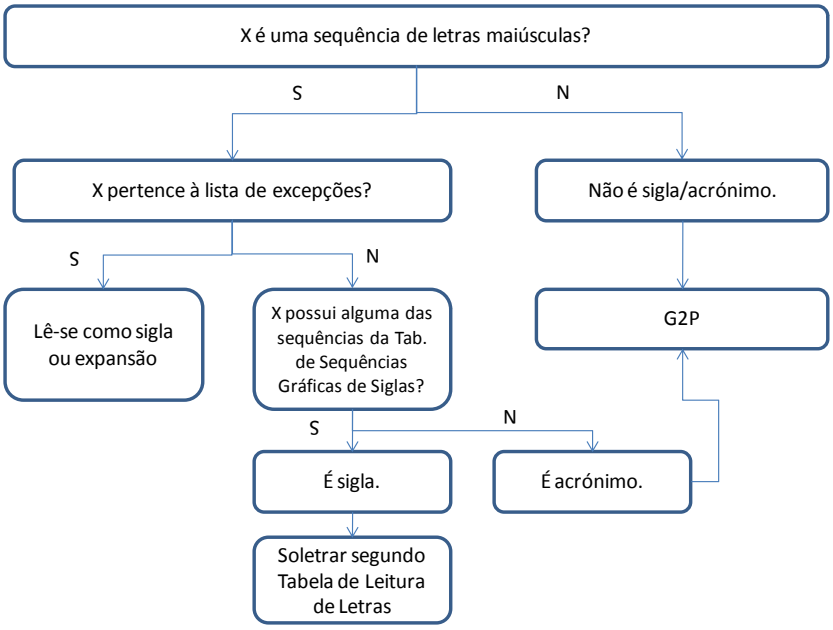

Figura 2: Algoritmo de leitura de siglas e acrónimos.

Em seguida, o algoritmo começa por verificar se uma dada palavra x é um candidato a sigla/acrónimo, pela presença ou não de sequências de letras maiúsculas. Se a resposta for positiva, estamos perante um candidato a sigla/acrónimo; se, pelo contrário, a resposta for negativa, segue para o algoritmo seguinte de transcrição grafema-fone (G2P). Na sequência do resultado positivo, o algoritmo vai verificar se a palavra x consta da Lista de Excepções (Tabela 4). Nesta tabela, constam também os casos em que a sigla é de origem estrangeira e é lida segundo o sistema fonológico de origem (ex. <MIT - Massachusetts Institute of Technology>).

Tabela 4: Lista de excepções do algoritmo de leitura de siglas/acrónimos.

Excepção	Transcrição fonética
AEP	a1. E1.pe1
AIP	a1.i1.pe1
APACDM	a1.pe1.a1.ce1.de1.E1.m@
CEE	se1.E1.E1
EUA	@S.ta1.du.zu.ni1.duZ.da.mE1.ri.k6
F1	fO1r.mu.l6.u~1
FAQ	Fa1.k@1
HIV	6.ga1.i1.ve1
IPE	i1.pe1.E1
IPO	i1.pe1.O1
MIT	E1.m@.a1j.ti1

Tabela 4: Lista de excepções do algoritmo de leitura de siglas/acrónimos (continuação).

MNE[100]	E1.m@.E1.n@.E1
OUA	O1.u1.1a
PAIGC	pe1.a1.i1.Ze1.se1
PE	pe1.E1
SA	E1.s@1.a1
SOS	E1.s@.O1.E1.s@
UA	u1.a1
UE	u1.E1
USE	u1.E1.s@.E1
ZEE	ze1.E1.E1

Na Tabela 4, encontram-se as palavras que apresentam os padrões gráficos de um acrónimo, tradicionalmente VCV (ex. <IPE – Investimentos e Participações Empresariais>), CVC (<HIV - Human Immunodeficiency Virus>) e que, por isso, se esperaria que fossem lidas, mas que, por razões desconhecidas, são antes soletradas, como as siglas. Existem casos ainda em que a palavra não se lê nem como sigla nem acrónimo, mas sim com a sua expansão, como se fosse uma abreviatura (ex. <F1 – Fórmula Um>, <EUA – Estados Unidos (da América)>). Caso uma dada palavra x não pertença à lista de excepções, é verificada a sua sequência gráfica. Se a palavra apresentar qualquer das sequências gráficas previstas na Tabela de Sequências Gráficas de Siglas (Tabela 5), então a palavra é soletrada segundo a transcrição fonética apresentada na Tabela de Leitura de Letras (Tabela 6). Caso contrário, é acrónimo, logo é transferida para o G2P. Em relação à Tabela 6, note-se ainda que o grafema <g> se lê [Ze1] apesar de a letra ser [ge1] (ex. <CGD>, <UGT>, <TGV>).

Tabela 5: Lista de sequências gráficas de siglas.

Sequência de grafemas	Exemplo
VC	AR, EP, UT
CC-[101]	BP, BT, BN, CD
CCC	BTT, PSD, CTT
CCCC	CGTP
VVVV	AIEA
VCC[102]	APL, ACP, ADN, EDP

[100] "Os acrónimos com a estrutura VCV que terminam em «E» são, muitas vezes, soletrados (IPE - Investimentos e Participações do Estado; USE - União dos Sindicatos de Évora), uma vez que esta vogal final pode ser ou não realizada. Por outro lado, e tal como acontece com os acrónimos terminados em «O» (IPO – Instituto Português de Oncologia), a sua leitura é passível de suscitar ambiguidade, não permitindo assim o reconhecimento ortográfico da sigla. Esta explicação não se aplica, contudo, a acrónimos como INE (Instituto Nacional de Estatística) e ESO (Observatório Europeu do Sul)." (Mendes *et al*. 2004)

[101] "Independentemente da sua extensão, são soletrados todos os acrónimos que incluem uma letra repetida, no início (CCAMB- Caixa de Crédito Agrícola Mútuo da Batalha, CCO-Centro de Coordenação Operacional)." (Mendes *et al*., 2004: 14)

Tabela 5: Lista de sequências gráficas de siglas (continuação).

VCCC[103]	ANMP, ANTT
CCV	BNU, BSE, MBA, SPA[104]
VCCV	ADSE, OCDE
CCCV	MPLA
C,V≥5[105]	CNLCS, ACCCIA

Todas as tabelas podem ser expandidas, robustecendo assim a performance do sistema. Os seguintes algoritmos foram implementados e testados e os seus resultados discutidos mais adiante neste capítulo.

Tabela 6: Tabela de leitura de letras.

Letra	Transcrição fonética	Letra	Transcrição fonética
A	a1	N	E1.n@
B	be1	O	O1
C	se1	P	pe1
D	de1	Q	ke1
E	E1	R	E1.R@
F	E1.f@	S	E1.s@
G	Ze1	T	te1
H	6.ga1	U	u1
I	i1	V	ve1
J	ZO.t6	W	d61.blju
K	ka1.p6	X	xi1S
L	E1.l@	Y	i1p.slOn
M	E1.m@	Z	ze1

2.6. Conversor de numerais

O problema da conversão de numerais é talvez o mais complexo deste módulo, dado que envolve a transcrição de números de telefone, datas, horas, números de

---

[102] "São igualmente soletrados todos os acrónimos com uma única vogal inicial (INCM – Imprensa Nacional-Casa da Moeda e AMVDN- Associação de Municípios do Vale Douro Norte)." (Mendes *et al*., 2004: 14)

[103] Ver nota de rodapé anterior.

[104] SPA é sigla para *Sociedade Portuguesa de Autores*. Existe também um homógrafo que se lê como acrónimo <SPA - Salutem per Acquam>. A desambiguação destes homógrafos necessita de informação contextual, pelo que será tratada em trabalho futuro.

[105] Sequências gráficas constituídas por consoantes ou vogais em número igual ou maior que 5 são soletradas, tal como defendem Mendes *et al*. (2004): "também no caso da Língua Portuguesa são vários, no nosso corpus, os exemplos de acrónimos não lidos, com tamanho igual ou superior a 5 letras: CNLCS (Comissão Nacional de Luta Contra a Sida), UIPSS (União das Instituições Particulares de Solidariedade Social), ACCCIA (Alto Comissariado Contra a Corrupção e a Ilegalidade Administrativa)."

conta, números romanos, numerais ordinais e, num nível mais sofisticado, expressões matemáticas[106] e fórmulas químicas. Além disso, muitas vezes surgem representações diferentes, por exemplo para horas (ex. <10:30 p.m.>, <22:30>), sendo necessário optar por uma leitura adequada.

Há, ainda, várias considerações a ter ao nível até da prosódia dos números de telefone. É comum fazerem-se pausas entre cada grupo de 3 dígitos nos números de telefone de Portugal (ex. +351 996 339 410). Por vezes essas pausas aparecem identificadas por espaços. Outras vezes, tal não acontece. No Brasil, por exemplo, essas pausas surgem identificadas por hífens depois dos prefixos de país e de estado (+55-11-9935-5539), mas tal pode também não acontecer.

Um dos principais problemas tem a ver com a dificuldade em identificar os números de telefone como tal e não como um número cardinal. Outros problemas surgem na leitura das seguintes expressões: <século XIV> e <XIV Festival Internacional de Teatro de Expressão Ibérica>, em que o número romano <XIV> se lê no primeiro caso como numeral cardinal <catorze> e no segundo caso como numeral ordinal <décimo quarto>.

Identificaram-se 3 grupos de numerais: 1) números árabes cardinais; 2) números árabes ordinais e 3) números romanos. Para cada um destes grupos foi elaborada uma tabela de conversão fonética e uma tabela de regras, como se passará a descrever nas secções seguintes.

2.6.1. Conversão de números árabes cardinais

A conversão de números árabes cardinais é um problema trivial e sobejamente tratado por muitas empresas, sobretudo bancos, departamentos financeiros, etc., para facilitar a impressão de cheques, notas de encomenda, ou quaisquer outros documentos em que os números árabes devam figurar por extenso. Este problema é, por exemplo, um exercício comum que se costuma colocar em níveis iniciais de cursos de programação. No entanto, não são conhecidos trabalhos publicados detalhando estes algoritmos de conversão, essenciais nesta fase de pré-processamento do texto.

Os algoritmos propostos nesta dissertação tratam separadamente as unidades, as dezenas, as centenas, os milhares e as dezenas de milhares, sendo que o resultado final é constituído pela soma do output de cada algoritmo. Convencionámos que o nosso sistema lê números entre 0 e 99 999. Os algoritmos fazem perguntas em função da posição do algarismo (unidade, dezena, centena, milhar, dezena de milhar) e consultam tabelas de transcrição (Tabelas 7, 8, 9 e 10) para obter o output.

---

[106] Não é nosso objectivo na presente dissertação contemplar a leitura de fórmulas matemáticas. No entanto, devemos destacar o trabalho que tem sido desenvolvido por Helder Ferreira ao longo de vários anos, no projecto AudioMath, disponível em: http://lpf-esi.fe.up.pt/~audiomath/pub_en.html#p2003 (28-12-2007) dedicado à leitura de fórmulas matemáticas para a comunidade de cegos e amblíopes e cujos pormenores podem ser consultados em Ferreira e Freitas (2004, 2005) e Ferreira (2005).

Tabela 7: Tabela de transcrição fonética de números árabes cardinais - unidades.

Número	Expansão	Transcrição Fonética
0	zero	zE1.ru
1	um	u~1
2	dois	do1jS
3	três	tre1S
4	quatro	kwa1.tru
5	cinco	si~1.ku
6	seis	s61jS
7	sete	sE1.t@
8	oito	o1j.tu
9	nove	nO1.v@

Tabela 8: Tabela de transcrição fonética de números árabes cardinais - 10-19.

Número	Expansão	Transcrição Fonética
10	dez	dE1S
11	onze	o~1.z@
12	doze	do1.z@
13	treze	tre1.z@
14	catorze	k6.to1r.z@
15	quinze	ki~1.z@
16	dezasseis	d@.z6.s61jS
17	dezassete	d@.z6.sE1.t@
18	dezoito	d@.zo1j.tu
19	dezanove	d@.z6.nO1.v@

Tabela 9: Tabela de transcrição fonética de números árabes cardinais - dezenas.

Número	Expansão	Transcrição Fonética
20	vinte	vi1~.t@
30	trinta	tri1~.t6
40	quarenta	kw6.re~1.t6
50	cinquenta	si~.kwe~1.t6
60	sessenta	s@.se~1.t6
70	setenta	s@.te~1.t6
80	oitenta	oj.te~1.t6
90	noventa	nu.ve~1.t6

Apresentam-se em seguida os algoritmos propostos para a conversão de números árabes cardinais. A expansão das abreviaturas usadas nas Figuras 3-7 é a seguinte: U (Unidades), D (Dezenas), C (Centenas), M (Milhares), DM (Dezenas de Milhares). Nos losangos estão representadas as perguntas que o algoritmo faz ao candidato a número árabe e nos quadrados apresentam-se as saídas do algoritmo, caso a resposta seja afirmativa (S) ou negativa (N). A principal dificuldade na construção destes algoritmos prendeu-se com a inclusão ou não inclusão da conjunção coordenada copulativa <e> entre os algarismos dos milhares e os algarismos das centenas, já que obedece a regras específicas (ex. <1_e001>, <1_e010>, <1_e099>, <1_e100> vs <1101>, <1876>, <1999>).

Tabela 10: Tabela de transcrição fonética de números árabes cardinais - centenas.

Número	Expansão	Transcrição Fonética
100	cem	s6~1j~
200	duzentos	du.ze~1.tuS
300	trezentos	tr@.ze~1.tuS
400	quatrocentos	kwa.tru.se~1.tuS
500	quinhentos	ki.Je~1.tuS
600	seiscentos	s6jS.se~1.tuS
700	setecentos	sE.t@.se~1.tuS
800	oitocentos	Oj.tu.se~1.tuS
900	novecentos	nO.v@.se~1.tuS

Estes algoritmos foram implementados em C/C++ e testados. A conversão de números árabes cardinais está na base da leitura de horas, datas, valores monetários, medidas, entre muitas outras ocorrências que surgem nos textos.

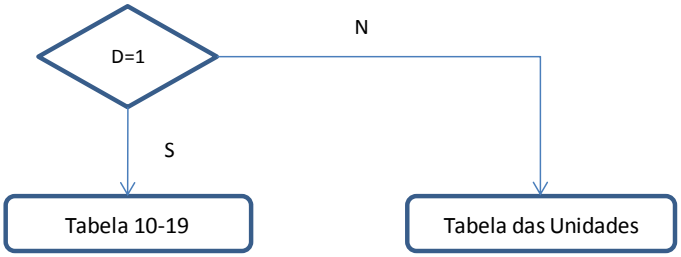

Figura 3: Algoritmo de conversão de números árabes cardinais - unidades.

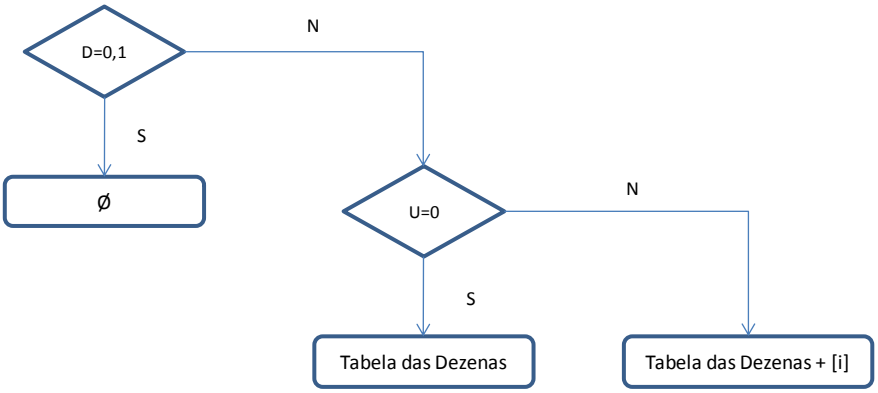

Figura 4: Algoritmo de conversão de números árabes cardinais - dezenas.

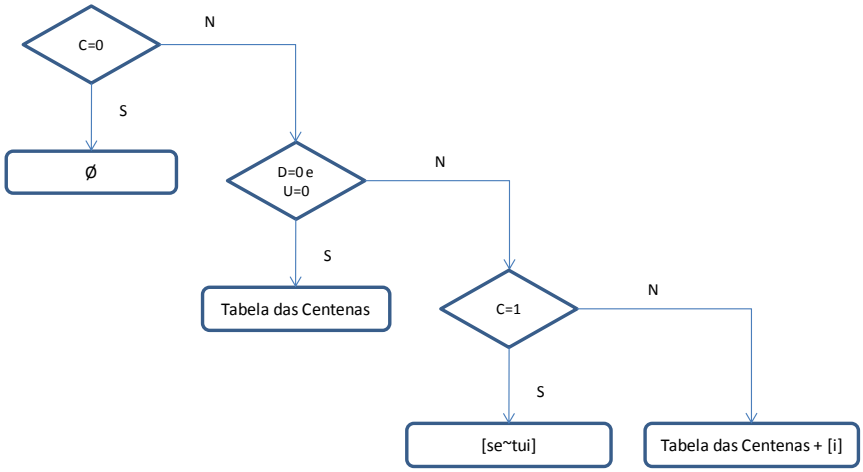

Figura 5: Algoritmo de conversão de números árabes cardinais - centenas.

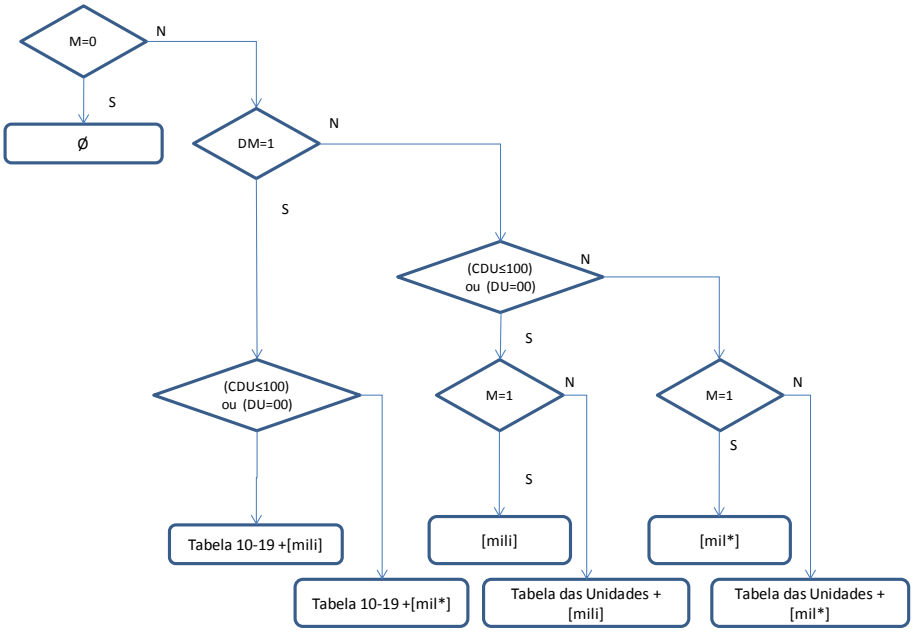

Figura 6: Algoritmo de conversão de números árabes cardinais - milhares.

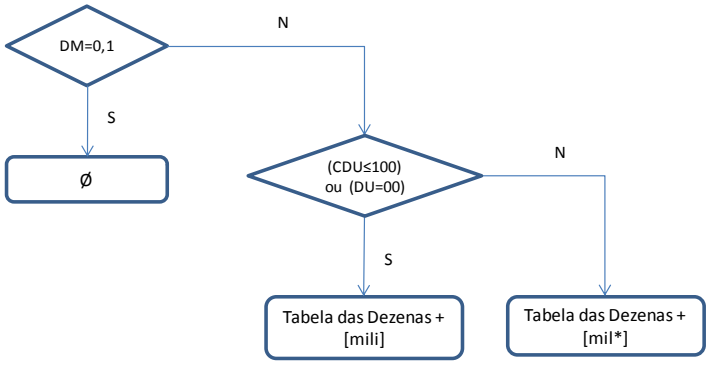

Figura 7: Algoritmo de conversão de números árabes cardinais – dezenas de milhares.

Para aplicações nas quais é possível ocorrer números superiores a dezenas de milhares, a extensão é natural.

2.6.2. Conversão de números árabes ordinais

A construção dos algoritmos de conversão de números árabes ordinais baseou-se na estrutura dos algoritmos anteriores propostos na secção 2.6.1. Apresentam-se algoritmos separados para o tratamento das unidades, das dezenas e das centenas (Figuras 8, 9 e 10). Os numerais ordinais apresentam uma formação bastante regular a partir das dezenas, o que facilita a sua conversão. Ao contrário dos numerais cardinais, apresentam flexão em género e número, embora a representação da flexão em género seja mais frequente, justificando o expoente <ª>. O algoritmo começa por procurar números árabes cardinais e em seguida verificar a presença dos caracteres <º> e <ª> em expoente. Se estas duas condições estiverem satisfeitas, o algoritmo faz a conversão das unidades, das dezenas e das centenas respectivamente (Tabelas 11, 12 e 13). A terminação [u], [6] da transcrição fonética é determinada pelo carácter em expoente <º>, <ª> respectivamente. O resultado final é, à semelhança do conversor anterior, a soma dos outputs das unidades, das dezenas e das centenas, respectivamente, e caso se verifiquem.

Tabela 11: Tabela de transcrição fonética de números árabes ordinais - unidades.

Unidades	Expansão	Transcrição Fonética
1º,ª	Primeiro(a)	pri.m61j.r(u,6)
2º,ª	Segundo(a)	s@.gu~1.d(u,6)
3º,ª	Terceiro(a)	t@r.s61j.r(u,6)
4º,ª	Quarto(a)	kwa1r.t(u,6)
5 º,ª	Quinto(a)	ki~1.t(u,6)
6 º,ª	Sexto(a)	s61jS.t(u,6)
7 º,ª	Sétimo(a)	sE1.ti.m(u,6)
8 º,ª	Oitavo(a)	oj.ta1.v(u,6)
9 º,ª	Nono(a)	no1.n(u,6)

Tabela 12: Tabela de transcrição fonética de números árabes ordinais - dezenas.

Dezenas	Expansão	Transcrição Fonética
10º,ª	Décimo(a)	dE1.si.m(u,6)
20º,ª	Vigésimo(a)	vi.ZE1.zi.m(u,6)
30º,ª	Trigésimo(a)	tri.ZE1.zi.m(u,6)
40º,ª	Quadrigésimo(a)	kw6.dr6.ZE1.zi.m(u,6)
50º,ª	Quinquagésimo(a)	ki~.kw6.ZE1.zi.m(u,6)
60º,ª	Sexagésimo(a)	sE.ks6.ZE1.zi.m(u,6)
70º,ª	Septuagésimo(a)	sE.ptw6.ZE1.zi.m(u,6)
80º,ª	Octogésimo(a)	O.ktO.ZE1.zi.m(u,6)
90º,ª	Nonagésimo(a)	no.n6.ZE1.zi.m(u,6)

Tabela 13: Tabela de transcrição fonética de números árabes ordinais - centenas.

Centenas	Expansão	Transcrição Fonética
100º,ª	Centésimo(a)	se~.tE1.zi.mu(u,6)
200º,ª	Ducentésimo(a)	du.se~.tE1.zi.m(u,6)
300º,ª	Tricentézimo(a)	tri.se~.tE1.zi.m(u,6)
400º,ª	Quadrigentézimo(a)	kw6.dri.Ze~.tE1.zi.m(u,6)
500º,ª	Quingentésimo(a)	ki~.Ze~.tE1.zi.m(u,6)
600º,ª	Seiscentésimo(a)	s6jS.se~.tE1.zi.m(u,6)
700º,ª	Septigentésimo(a)	sE.pti.Ze~.tE1.zi.m(u,6)
800º,ª	Octogentésimo(a)	O.kti.Ze~.tE1.zi.m(u,6)
900º,ª	Nogentésimo(a)	no~.Ze~.tE1.zi.m(u,6)

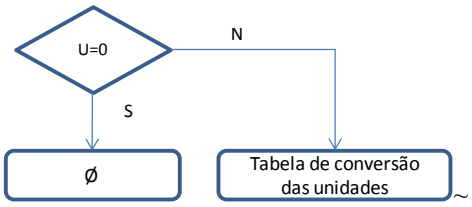

Figura 8: Algoritmo de conversão dos números árabes ordinais: unidades (U).

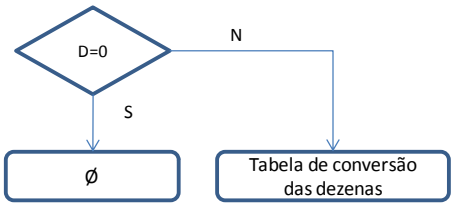

Figura 9: Algoritmo de conversão dos números árabes ordinais: dezenas (D).

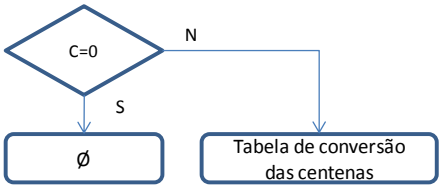

Figura 10: Algoritmo de conversão dos números árabes ordinais: centenas (C).

2.6.3. Conversão de números romanos

A conversão de números romanos é outro problema que merece alguma atenção. A sua utilização surge associada sobretudo a títulos monárquicos (ex. <D João I>, <Carlos V>), à numeração dos séculos (ex. <século IV>, <século XVI>) e à numeração de certos eventos, feiras, certames, congressos periódicos (ex. <XXIII Encontro Nacional da Associação Portuguesa de Linguística>). Na verdade, os números romanos são sempre convertidos para o extenso dos números árabes, até porque a designação dos números provém do étimo latino, como aliás toda a Língua Portuguesa. A questão aqui não reside tanto na sua conversão mas sim no tipo de conversão. É que a conversão de números romanos assume ora um extenso numérico cardinal (ex. <século XV> [sE1.ku.lu.ki~1.z@]) ora um extenso numérico ordinal: (ex. <Afonso X> [6.fo~1.su.dE1.si.mu]). E apesar de mais uma vez se tratar de um problema trivial, não se encontram algoritmos publicados para a conversão de números romanos.

Para fazer face a estas questões, foram propostas uma tabela de conversão de números romanos para árabes (Tabela 14) e uma tabela de regras de conversão de números romanos para cardinais ou ordinais (Tabela 15). Na Tabela 16, apresentam-se alguns casos em que o candidato a número romano é na verdade uma letra e, por isso, é lido segundo a Tabela 6. Considerámos apenas os números romanos até cinquenta, por serem mais frequentes, embora esta tabela possa ser aumentada em qualquer momento. O sistema começa por fazer a conversão do número romano. Em seguida, aplica-lhe as regras.

Tabela 14: Tabela de conversão de números romanos.

Número romano (NR)	Conversão número árabe cardinal (NAC)	Conversão número árabe ordinal (NAO)
I	1	1º
II	2	2º
III	3	3º
IV	4	4º
V	5	5º
VI	6	6º
VII	7	7º
VIII	8	8º
IX	9	9º
X	10	10º

Tabela 14: Tabela de conversão de números romanos (continuação).

XI	11	11º
XII	12	12º
XIII	13	13º
XIV	14	14º
XV	15	15º
XVI	16	16º
XVII	17	17º
XVIII	18	18º
XIX	19	19º
XX	20	20º
XXI	21	21º
XXII	22	22º
XXIII	23	23º
XXIV	24	24
XXV	25	25º
XXVI	26	26º
XXVII	27	27º
XXVIII	28	28º
XXIX	29	29º
XXX	30	30º
XXXI	31	31º
XXXII	32	32º
XXXIII	33	33º
XXXIV	34	34º
XXXV	35	35º
XXXVI	36	36º
XXXVII	37	37º
XXXVIII	38	38º
XXXIX	39	39º
XL	40	40º
XLI	41	41º
XLII	42	42º
XLIII	43	43º
XLIV	44	44º
XLV	45	45º
XLVI	46	46º
XLVII	47	47º
XLVIII	48	48º
XLIX	49	49º
L	50	50º

A previsão da flexão em género do número romano ao ser convertido para número árabe cardinal necessita, neste caso, de um analisador morfológico com informação de género, o que não foi contemplado nesta fase do trabalho, mas que está previsto como trabalho futuro.

Tabela 15: Regras de conversão de números romanos (NR).

#	Regra	Exemplo
1	Tabela 16 de Excepções	
2	Se NR ≥ 11 e (P-1[107] = P_M[108] ou <século>, <séc.>) → NAC	<século XI>, <João XXI>
3	Se NR ≤ 10 e (P-1 ou P+1 = P_M ou <século>, <séc.>) → NAO	<D. João IV>, <Manuel I>,<XX Festival Internacional de Teatro Ibérico>, <I Divisão>
4	Se <I>, não consta em nenhum contexto anterior → [i1]	<lista I>
5	Se <V> não consta em nenhum contexto anterior → [ve1]	<lista V>
6	Se <X> não consta em nenhum contexto anterior → [Si1S]	<tenho X hectares de terra>
7	Se <L> não consta em nenhum contexto anterior → [E1.l@]	<Lumière L>

Tabela 16: Tabela de excepções na conversão de números romanos (NR).

Grafema <I>	**Transcrição fonética**
<I love you>[109]	[a1j]
Grafema <V>	**Transcrição fonética**
<16 V>	[va1l*vul6S]
Grafema <X>	**Transcrição fonética**
<cromossoma X> <Malcolm X> <triplo X> <raio(s) X> <geração X> <senhor X>	[Si1S]
Grafema <L>	**Transcrição fonética**
<l‘>	[l]

2.6.4. Leitura de horas, datas e números com casas decimais

Para estes casos, fez-se uma listagem das várias possibilidades de ocorrência gráfica e em seguida estabeleceram-se regras de conversão. Apesar de este assunto surgir depois da conversão de numerais por uma questão de organização do trabalho, a nível de implementação é processado antes da conversão de numerais.

[107] Designa palavra anterior ao número romano.
[108] Designa palavra começada por maiúscula.
[109] O registo desta excepção explica-se apenas porque no corpus de análise utilizado do Cetem-Público apresentava uma relativa frequência de ocorrência.

Tabela 17: Regras para leitura de datas.

#	Leitura de datas	Exemplo
1	Se NAC ≤31 + "." ou "/" ou "-"+ NAC ≤12 + "." ou "/" ou "-"+NAC ≤ 9999 → TNAC [du] TNAC [d@] TNAC[110]	1.8.2007, 01-08-2007, 1/8/2007 → [u~1.du.o1j.tu.d@.do1jS.mi.li.sE1.t@]

A leitura de horas necessitou de uma regra prévia de concordância em género do número árabe cardinal 1 e 2 com o substantivo <hora>, justificando assim as regras 1 e 2 da leitura de horas.

Tabela 18: Regras para leitura de horas.

#	Leitura de horas	Exemplo
1	Se algarismo das unidades do NCA = 1 → [um6]	1h → [u1.m6.O1.r6]
2	Se algarismo das unidades do NCA = 2→ [du6S]	22h→ [vi~1.t@.i.du1.6.zO1.r6S]
3	Se NAC + "h" + SP, Pont → TNCA + [Or6S]	15h → [ki~1.z@.O1.r6S]
4	Se NAC + "h" ou ":" + NAC → TNCA + [i] + TNCA	22:30, 22h30 → [vi~1.t@.i.du1.6.zi.tri~1.t6]
5	Se NAC + <h>+NAC+<m> → TNCA + [O1.r6S.i] + TNCA [mi.nu1.tuS]	10h 15m→ [dE1.zO1.r6S.i.ki~1.z@. mi.nu1.tuS]

Tabela 19: Regras para leitura de números com casas decimais.

#	Leitura de números com casas decimais	Exemplo
1	Se NCA + "," + NCA→ TNCA + [virgul6] + TNCA	12,5% → [do1.z@.vir1.gu.l6.si~1.ku.pur.se1~.tu]
2	Se NCA + "." + NCA→ TNCA + [po~tu] + TNCA	3.6% →[tre1S.po~1.tu.´s61jS.pur.'se~1.tu]

O sistema começa por procurar algarismos. Em seguida, procura padrões e aplica-lhe regras (Tabela 17, 18, 19). Apesar de haver outros sistemas de representação de datas e horas, como o sistema americano (ex. 12/29/06) e de haver várias possibilidades de leitura mesmo em português para datas (ex. <1.8.2007> pode ser lido

[110] Descrição da regra 1 - Perante a seguinte ocorrência: número árabe cardinal (NAC) menor ou igual a 31, seguido de ponto final ou barra à direita ou hífen, seguido de número árabe cardinal menor ou igual a 12 seguido de ponto final ou barra à direita ou hífen, seguido de número árabe cardinal menor ou igual a 9999, trata-se de uma data e a sua leitura deve ser a respectiva transcrição fonética do número árabe cardinal (TNAC), seguida de [du], seguida da respectiva transcrição fonética do número árabe cardinal (TNAC) que se encontra em segunda posição, seguida de [d@], seguida da respectiva transcrição fonética do número árabe cardinal (TNAC) que se encontra em terceira posição.

como “um do oito de dois mil e sete” ou “um de Agosto de dois mil e sete”) e horas (ex. <22h30> pode ser lido “vinte e duas horas e trinta minutos”, “dez e meia da noite” ou “vinte e duas e trinta”), optou-se apenas por uma leitura possível, por razões de simplificação da programação.

Com estes casos, ficam assim cobertas várias ocorrências especiais de numerais em português. Os testes da performance do módulo de pré-processamento serão apresentados em seguida.

Tabela 20: Regras para outros casos.

#	Pontuação desportiva	Exemplo
1	Se NCA+”-“+NAC → TNAC +[6]+TNAC	<2-1> [do1jS6u~1]
#	**Medidas**	**Exemplo**
1	Se NAC + “x” +NAC → TNAC [pur] TNAC	<10x15> [dE1S.pur.ki~1.z@]

2.7. Testes e discussão dos resultados

Todos os sub-módulos apresentados nas secções anteriores foram implementados e testados. Apresentaremos nesta secção os testes relativos ao desempenho do leitor de siglas e acrónimos e do conversor de numerais, incluindo a conversão de horas, datas e números com casas decimais, uma vez que são os únicos sub-módulos em que existem regras e em que os resultados não são triviais. Os restantes sub-módulos, nomeadamente os separadores de frase e de palavra, o conversor de símbolos e o expansor de abreviaturas não foram testados para além das listas apresentadas, porque são de implementação trivial e com conversão/expansão directa.

Foram reunidos dois tipos de corpora extraídas do Cetem-Público através de um algoritmo de busca e extracção de entidades desenvolvido por Luís Coelho: um com 217 frases, 7248 palavras, contendo 249 siglas, acrónimos e números romanos e outro *corpus* com 185 frases e 5755 palavras, contendo dígitos. O primeiro *corpus* tinha como objectivo testar o leitor de siglas, acrónimos e números romanos; o segundo foi desenhado para testar o conversor de numerais árabes, horas, datas, números com casas decimais e outros casos (pontuação desportiva e medidas).

Na Tabela 21, apresentam-se os resultados dos testes não só do conversor de siglas e acrónimos, mas também do conversor de numerais romanos, visto que em ambos os casos se trata da conversão de letras maiúsculas. Da análise dos resultados obteve-se 0,12% de erros, sendo a sua totalidade decorrente da incorrecta leitura de siglas como acrónimos. Das 169 siglas presentes no *corpus*, 19 são palavras que pertencem à Tabela 4 de excepções (ex. <EUA>, que aparece 15 vezes e <CEE>, com 4 ocorrências) e 9 são assim erros, siglas cujas sequências não constam da Tabela 5 (ex. <MOPTC>, <VVER>, <ASGFIM>, <IIPA>, que aparece 4 vezes, <SEADR> e <PAICV>). Do conjunto de 20 candidatos a números romanos, ocorreram 15 conversões para numerais árabes ordinais, 2 conversões para numerais árabes cardinais e 3 excepções, contidas na Tabela 16. Estes resultados andam muito

próximos de avaliações de outros sistemas análogos para o português (98,05% em Barbosa *et al.*, 2003b para o PB e 99,4% em Trancoso & Viana, 1997, para o PE).

Tabela 21: Resultados do teste do conversor de siglas, acrónimos e números romanos.

Elemento testado	# corpus	% corpus	# erro	% erro
Siglas	169	2,33	9	0,12
Acrónimos	60	0,83	0	0,0
Romanos	20	0,28	0	0,0
Total	**249**	**3,44**	**9**	**0,12**

Na Tabela 22, mostram-se os resultados do teste do conversor de numerais. De um conjunto de 289 dígitos analisados, representanto 5,02% do *corpus* extraído, o sistema errou 8 vezes, representanto 0,14% de erros por palavra. Os erros ocorreram na leitura de cronometragem em provas envolvendo tempo, em casos como <1m29,246s> e em representações de distâncias segundo medidas internacionais <10.000 m> (o sistema interpreta este dígito como um número com casas decimais, mas o contexto desambigua para <dez mil metros>).

Tabela 22: Resultados do teste do conversor de numerais.

Elemento testado	# corpus	% corpus	# erro	% erro
NAC	200	3,48	0	0,0
NAO	15	0,26	0	0,0
Datas	2	0,03	0	0,0
Horas	12	0,21	0	0,0
Nºs c/ casas decimais	41	0,71	0	0,0
Outros	19	0,33	8	0,14
Total	**289**	**5,02**	**8**	**0,14**

Como trabalho futuro, implementaremos regras para a leitura de números de telefone, endereços de e-mail, urls e endereços de internet.

2.8. Aplicações do sistema ao português do Brasil

A grande proximidade linguística entre o PE e o PB permite fazer adaptações muito imediatas nos sub-módulos apresentados.

A separação de frase e de palavra aplica-se directamente com excepção da lista de abreviaturas com ponto, que para o PB foi revista. A lista de abreviaturas com ponto para o PB é assim constituída por 109 títulos (ex. <col.> → <coronel> e formas de tratamento (ex. <Exa.> →<Excelência>) e por cerca de 3000 abreviaturas comuns (ex. <subj.> → <subjuntivo>)[111].

[111] Fonte: http://pt.wikipedia.org/wiki/Lista_de_abreviaturas (28-12-2007).

A conversão de símbolos e caracteres especiais em PB é a mesma que em PE, com as seguintes diferenças na designação do caracteres: <#> (jogo da velha), <©> (marca registrada), <¥> (iene). A expansão de abreviaturas é realizada através da lista de abreviaturas com ponto utilizada no separador de frases. Essas expansões de símbolos e de abreviaturas serão depois transcritas foneticamente pelo conversor grafema-fone para PB (vide capítulo 5).

Em relação ao leitor de siglas e acrónimos, e apesar de haver diferenças entre PE e PB, as regras permanecem as mesmas, dado que a leitura de siglas ou acrónimos decorre da estrutura silábica da língua, que apresenta essencialmente a mesma estrutura em ambas as variedades do português (Mateus & Andrade, 2000: 38-64). As diferenças residem na lista de excepções, que teria de ser adaptada ao PB e na tabela de leitura de letras, que difere do PE na transcrição fonética das seguintes: <f> [E1.fi], <g> [Ze1], <k> [ka1], <l> [E1.li], <m> [E1.mi], <n> [E1.ni], <r> [E1.Ri], <s> [E1.si], <w> [da1.bliw] e <y> [i1.psi.lo~]. Em trabalho futuro, testaremos o leitor de siglas e acrónimos aqui descrito com textos reais.

A conversão de numerais árabes e romanos em PB funciona de maneira análoga à do PE. As diferenças residem na ortografia dos seguintes numerais: <16> (dezesseis), <17> (dezessete), <19> (dezenove), <50> (cinqüenta) e <50º/50ª> (qüinquagésimo/a) e na transcrição fonética das tabelas de expansão de numerais[112]. Será necessário que neste ponto actue um conversor grafema-fone para PB, como será descrito no capítulo 5. Testes da conversão de numerais utilizando textos brasileiros serão também trabalho futuro.

2.9. Aplicações do sistema ao galego

À semelhança do que acontece com o PB, é possível aplicar o módulo de normalização de texto apresentado neste capítulo ao galego, com poucas adaptações.

Também os separadores de frase e de palavra se aplicam directamente, visto que o galego apresenta regras de utilização de pontuação e de hífen semelhantes ao português, excepto, uma vez mais, a lista de abreviaturas com ponto, que deverá ser revista e adaptada ao galego.

A conversão de símbolos e caracteres especiais em galego não é muito diferente do PE (veja-se Tabela 23), podendo ser facilmente adaptada no nosso programa.

Tabela 23: Símbolos e sua designação em galego.

Símbolo	Conversão ortográfica	Símbolo	Conversão ortográfica
#	cardinal	∞	infinito
$	dólar	μ	miu
%	por cento	α	alpha
&	e comercial	β	beta
*	asterisco	Γ, γ	gamma

[112] Além destes casos, <bilião> (mil milhões) em PE se diz <bilhão> em PB. Expansões desta magnitude não foram tratadas neste trabalho, mas serão previstas em trabalho futuro.

Tabela 23: Símbolos e sua designação em galego (continuação).

+	máis	Δ, δ	delta
-	menos	ε	épsilon
/	barra á dereita, ou	η	eta
=	igual a	ζ	zeta
@	arroba	θ	teta
\	barra á esquerda	ι	iota
_	underscore	Λ,λ	lambda
~	til	ν	niu
£	libra	ξ	xi
¥	yen	Π, π	pi
€	euro	ρ	ro
©	copyright	Σ, σ	sigma
®	marca rexistrada	τ	tau
° C	grao(s) Celsius	φ	phi
÷	a dividir por	χ	chi
×	veces	ψ	psi
≤	menor ou igual a	ω	omega
≥	maior ou igual a	^{3}	cúbico(s)
≠	diferente de	^{2}	cadrado(s)

Na Tabela 24, apresentamos algumas abreviaturas mais frequentes em galego. Esta tabela poderá ser expandida.

Tabela 24: Abreviaturas e sua expansão em galego.

Abreviatura	Conversão ortográfica	Abreviatura	Conversão ortográfica
a.	ano	fols.	folios
adic.	adición	Gal.	galego/a
admón.	administración	ibid.	ibidem
adv.	adverbio	Ilmo.	Ilustrísimo
adx.	adxectivo	it.	Italiano/a
al.	alemán/alemá	l.	Liña
ant.	antigo	ll.	liñas
antrop.	antropónimo	masc.	masculino
apdo.	apartado	n.v.	edición non venal
ar.	árabe	nº	número
art.	artigo	núms.	números
bras.	brasileiro	p.	páxina
c.	circa	pl.	plural
cap.	capítulo	port.	portugués/portuguesa
cast.	castelán/castelá	pp.	páxinas
cat.	catalán/catalana	prof.	profesor
col.	colección	pte.	presidente
coord.	coordinador/a	q.D.g.	que Deus garde
D.	don	S.	Santo
d.e.p.	descanse en paz	s.a.	ano de publicación sen especificar
Dª.	Dona	s.s.s.	o seu seguro servidor
doc.	documento	séc.	século
dr.	doutor	sg.	singular
dra.	doutora	ss.	seguintes
Ex.	exemplo	vde.	vostede
Excmo.	excelentísimo	V.V.A.A..	varios autores
fol.	folio		

Muito embora se esperem tendências semelhantes às enunciadas para o PE por Mendes *et al.* (2004), dada a proximidade da estrutura silábica do PE com o galego, impõe-se uma investigação mais aprofundada em relação à leitura de siglas e acrónimos em galego, já que ocorrem certas articulações inesperadas que não passam por soletrar a sigla, como acontece com o PE e PB:

> "Ao se referirem a nomes propios, a motivación é mais escura para o falante, quen en moitas ocasións se mostra incapaz de desenvolver o seu contido, de modo que intentará pronunciar as iniciais adaptándoas ás características fonéticas da lingua: PSOE = pesoe/soe, BNG = benegá, ILG = ilga." (Freixeiro, 2006b)

A Figura 11, extraída das Normas Ortográficas e Morfolóxicas do Idioma Galego (2003: 9), ilustra o nome e transcrição fonética das letras em galego.

LETRA	NOME	PRONUNCIA
a	a	[a]
b	be	[b]
c	ce	[θ] (ou [s]), [k]
d	de	[d]
e	e	[e], [ɛ]
f	efe	[f]
g	gue	[g] (ou [h])
h	hache	(cero)
i	i	[i]
l	ele	[l]
m	eme	[m]
n	ene	[n]
ñ	eñe	[ɲ]
o	o	[o], [ɔ]
p	pe	[p]
q	que	[k]
r	erre	[ɾ], [r]
s	ese	[s]
t	te	[t]
u	u	[u]
v	uve	[b]
x	xe	[ʃ], [ks]
z	zeta	[θ] (ou [s])

Figura 11: Alfabeto em galego.

A conversão de numerais em galego tem funcionamento análogo ao descrito para PE. As únicas adaptações dos algoritmos apresentados consistem na actualização da ortografia dos numerais, cardinais e ordinais (vejam-se Figuras 12 e 13, extraídas de *Normas Ortográficas e Morfolóxicas do Idioma Galego* (2003: 76-77), e pela respectiva actualização fonética. As formas compostas dos numerais ordinais em galego apenas admitem morfema de género feminino no segundo elemento (ex. <décimo primeiro/décimo primeira>), ao contrário do que acontece em português, em

que os dois elementos admitem forma feminina (ex. <décimo primeiro/décima primeira>).

cero	*vinte e dous / vinte e dúas*, etc.
un/unha	*trinta*
dous/dúas	*trinta e un, trinta e unha*, etc.
tres	*corenta*
catro	*cincuenta*
cinco	*sesenta*
seis	*setenta*
sete	*oitenta*
oito	*noventa*
nove	*cen*
dez	*cento un, cento unha*, etc.
once	*douscentos / duascentas*
doce	*trescentos / trescentas*
trece	*catrocentos / catrocentas*
catorce	*cincocentos / cincocentas,*
quince	*quiñentos/ -as*
dezaseis	*seiscentos / seiscentas*
dezasete	*setecentos / setecentas*
dezaoito	*oitocentos / oitocentas*
dezanove	*novecentos / novecentas*
vinte	*mil*
vinte e un / vinte e unha	

Figura 12: Numerais cardinais em galego.

primeiro	*décimo sétimo*
segundo	*décimo oitavo*
terceiro	*décimo noveno*
cuarto	*vixésimo*
quinto	*vixésimo primeiro*
sexto	*vixésimo segundo*, etc.
sétimo	*trixésimo*
oitavo	*cuadraxésimo*
noveno	*quincuaxésimo*
décimo	*sesaxésimo*
undécimo / décimo primeiro	*septuaxésimo*
duodécimo / décimo segundo	*octoxésimo*
décimo terceiro	*nonaxésimo*
décimo cuarto	*centésimo*
décimo quinto	*milésimo*
décimo sexto	*millonésimo*

Figura 13: Numerais ordinais em galego.

A conversão de numerais romanos passa primeiro por uma conversão para numerais árabes e em seguida para a sua expansão e conversão fonética. A conversão fonética dos numerais em galego é feita pelo transcritor grafema-fone apresentado no

capítulo 5. A leitura de horas e datas pode também ser adaptada ao galego, após as alterações ortográficas acima referidas.

Como se procurou demonstrar, são poucas as diferenças existentes ao nível do pré-processamento em galego, pelo que uma adaptação do módulo apresentado nesta dissertação é tarefa simples que exige apenas um estudo mais aprofundado ao nível da integração das siglas na estrutura silábica do galego actual.

2.10. Síntese do capítulo 2

As principais ideias-chave a reter no final deste capítulo são as seguintes:

- O pré-processamento ou normalização de texto é uma tarefa dependente da língua que envolve a conversão de toda a espécie de símbolos, abreviaturas, siglas, acrónimos, numerais, fórmulas matemáticas, endereços, números de telefone, etc. em sequências ortográficas que serão depois transcritas foneticamente;
- O pré-processamento apresentado neste trabalho é constituído por: separador do texto em frases, separador das frases em em palavras, conversor de símbolos e caracteres especiais, expansor de abreviaturas, leitor de siglas e acrónimos, conversor de numerais (árabes cardinais, árabes ordinais, romanos) e leitor de horas, datas e números com casas decimais;
- Este módulo foi implementado e testado com corpora reais, tendo-se obtido 99,88% de acerto para o conversor de siglas/acrónimos e números romanos e 99,86% de acerto para o conversor de dígitos (numerais árabes cardinais e ordinais, datas, horas, números com casas decimais, medidas e pontuação desportiva);
- A grande proximidade linguística entre o PE e o PB e entre o PE e o galego permite aplicar os sub-módulos apresentados com poucas adaptações envolvidas;
- Como trabalho futuro, implementaremos regras para a leitura de números de telefone, endereços de e-mail, urls e endereços de internet, faremos testes da conversão de numerais utilizando textos brasileiros e um estudo mais aprofundado ao nível da integração das siglas na estrutura silábica do galego actual.

Capítulo 3

Desambiguador de homógrafos

A desambiguação de homógrafos heterófonos constitui um dos problemas de mais difícil solução na conversão texto-fala. Em português, é responsável por 0,62% de taxa de erro, valor resultante do teste com o nosso *corpus* Cetem_Público de 1000 frases[113], o que significa que de entre 9090 palavras que compõem o *corpus*, há 57 homógrafos com problemas de transcrição. Neste capítulo, propomos dois tipos de soluções: 1) algoritmos baseados na análise morfossintáctica para desambiguar pares de homógrafos que pertencem a classes gramaticais diferentes; 2) algoritmos baseados na análise semântica para desambiguar pares de homógrafos que pertencem à mesma classe gramatical. Foram propostos 24 algoritmos baseados em regras linguísticas para solucionar um elenco de 116 pares de homógrafos. Foram conduzidos vários testes ao sistema, sendo a sua melhor performance de 98,2% de acerto. Os resultados foram discutidos, bem como a sua aplicabilidade ao português do Brasil e ao galego.

Este sistema permite dar resposta ao problema da leitura dos homógrafos na conversão texto-fala do português.

Versões prévias deste capítulo foram publicadas em artigos científicos com revisores e podem ser encontradas em:

- Braga, D.; Coelho, L.; Resende Jr., F.G.V. 2007. "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", *roceedings of Interspeech 2007*, 27-31 Agosto de 2007, Antuérpia, Bélgica.
- 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.

[113] Este *corpus* de teste encontra-se descrito no Capítulo 6. Trata-se de 1000 frases extraídas aleatoriamente do corpus jornalístico Cetem-Público, (disponível para consulta online em http://www.linguateca.pt/CETEMPublico/ - 18-12-2007) que serviu de *corpus* de teste dos vários módulos apresentados neste trabalho e que foi gravado em estúdio profissional, constituindo assim também a base de dados de voz que alimenta o motor de síntese (cf. Capítulo 6).

3.1. 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 poderão decidir se se deve ler a palavra com vogal tónica aberta ou fechada.

A ambiguidade dos homógrafos heterófonos, exemplificada em pares do tipo <o acerto> [e] 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[114] (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).

[114] 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.

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). 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 (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 de TTS, destacam-se as propostas de Ribeiro *et al.* (2002, 2003) para o PE, Seara *et al.* (2001, 2002) e Barbosa *et al.* (2003c) 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.

3.2. Arquitectura do desambiguador de homógrafos heterófonos

O desambiguador de homógrafos heterófonos constitui uma componente do módulo de Análise Fonética (vide Figura 1), articulando-se directamente com o Conversor Grafema-Fone[115]. 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.

Na Figura 14, pode ver-se a estrutura do desambiguador de homógrafos heterófonos. O desambiguador pode ser encarado como uma parte do analisador morfossintáctico. Na verdade, o seu funcionamento está dependente da análise morfossintáctica, por ser necessário identificar as categorias gramaticais das palavras que ocorrem à esquerda e à direita do homógrafo em análise.

A partir da observação da Figura 14, pode ver-se que o input do sistema é o texto que é separado em palavras. 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 24 tipos de homógrafos, que são apresentados na secção 3.3., o que faz com que haja 24 outputs possíveis nesta fase do sistema.

[115] Também designado por Conversor LTS (Letter-to-Sound) ou G2P (Grapheme-to-phone/me). Este módulo será descrito no Capítulo 5.

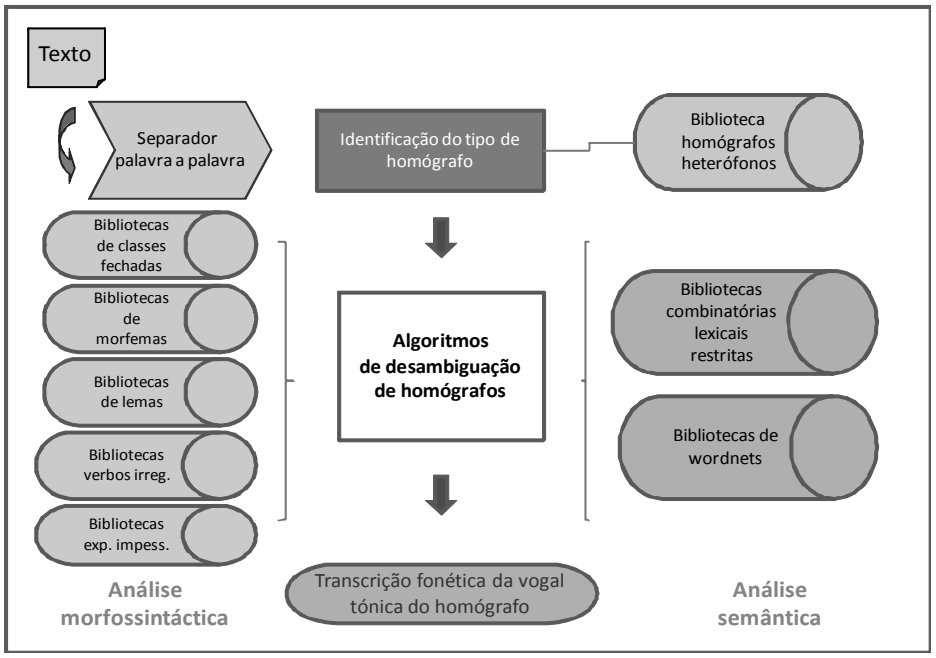

Figura 14: Arquitectura do desambiguador de homógrafos heterófonos.

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. 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:

    - preposições (PREP)
    - advérbios (ADV) e advérbios de quantidade (ADV_Q)
    - contracções de preposição com determinante/pronome (CONT)
    - conjunções subordinativas (C_S) e locuções conjuncionais subordinativas (Loc_S)

- conjunções coordenativas (C_C) e locuções conjuncionais coordenativas (Loc_C)
- determinantes artigos indefinidos (ART_IND)
- pronomes e determinantes demonstrativos (DEM)
- pronomes e determinantes possessivos (POSS)
- pronomes e determinantes indefinidos (IND)
- pronomes e determinantes interrogativos (INT)
- numerais (NUM)
- pronomes pessoais sujeito (P_PES_SU) e pronomes pessoais objecto (P_PES_O_1)[116], (P_PES_O_2)[117], (P_PES_O_3)[118]
- pronomes relativos (P_REL)
- interjeições (INTJ)

2) **Biblioteca de afixos**[119], constituída pelas seguintes sub-classes:

   - Sufixos Nominais (Des_N), Adjectivais (Des_Adj) e Adverbiais (Des_Adv)
   - Sufixos verbais (Des_V)
   - Prefixos portugueses (Pref_PT)
   - Radicais gregos e latinos (R_GL)

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

4) **Biblioteca de expressões unipessoais** (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[120],** constituída pelo dicionário Jspel[121] para o Português, com cerca de 34000 palavras, anotado morfologicamente, que resultou do

---

[116] Pronomes pessoais objecto que não sofram processos de assimilação resultantes da co-articulação com formas verbais (ex: <me>, <te>, <se>, <lhe>...).

[117] 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>).

[118] Pronomes pessoais objecto na terceira pessoa que sofrem assimilação no contacto com formas verbais com nasal ou ditongo em situação implosiva (ex: compram-no).

[119] 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." (Mateus et al., 2003: 941).

[120] 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

projecto Natura, ainda em curso[122], levado a cabo por investigadores do pólo de Braga da Linguateca[123], 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.* (2007). 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 116 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[124] e as bibliotecas de *Wordnets*, cujo funcionamento será descrito na secção 3.3.

---

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.

[121] 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 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, 18-12-2007).

[122] Disponível para download em: http://natura.di.uminho.pt/wiki/index.cgi?jspell (18-12-2007).

[123] Para mais informação sobre o pólo de Braga da Linguateca, consultar :http://linguateca.di.uminho.pt/ (18-12-2007).

[124] 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

As bibliotecas de combinatórias lexicais restritas abrangem, segundo a designação de Iriarte (2001), os frasemas completos (ou expressões idiomáticas)[125] (ex: "cor de burro quando foge".), os semi-frasemas (ou colocações)[126] (ex: "pregar um susto") e os quase-frasemas[127] (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[128], o COMPARA[129] (*corpus* paralelo em Português e em Inglês constituído por textos literários e suas traduções) e o EUROPARL – Opus[130] (constituído por transcrições dos debates do parlamento europeu; *corpus* alinhado para 12 línguas).

As bibliotecas de *wordnets*[131], obtidas pelo mesmo processo que as anteriores, procuram reunir palavras semanticamente relacionadas previsíveis de co-ocorrerem

---

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).

[125] 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 soma 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'.".

[126] 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 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).

[127] 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).

[128] Disponível em: http://www.linguateca.pt/CETEMPublico/ (18-12-2007).

[129] Disponível em: http://adamastor.linguateca.pt/COMPARA/BuscaSimples.html (18-12-2007).

[130] Disponível em: http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=pt (18-12-2007).

[131] 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/ (18-12-2007). Está em curso o projecto de criação de uma

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*.

3.3. Algoritmos de desambiguação de homógrafos heterófonos

3.3.1. 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 116 homógrafos (Tabelas 22 e 23) 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 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 para isso 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 seguidamente.

Wordnet para o português , no Centro de Linguística de Lisboa: http://www.clul.ul.pt/clg/projectos/WordNet.PT.html (18-12-2007), mas com resultados ainda não disponibilizados.

3.3.2. Tipologia de homógrafos heterófonos

Nas Tabelas 25 e 26 que se seguem, apresentam-se as tipologias de homógrafos consideradas. Na Tabela 25, 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 116 pares de homógrafos, 73 pertencem aos tipos 1 e 2, ou seja, 62,9% 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, 7, 8, 10, 12, 14, 15 e 16, embora os tipos 14, 15 e 16 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 as vogais do radical serem frequentemente semi-fechadas nas formas nominais, ao passo que nas formas verbais elas se tornam invariavelmente semi-abertas.

O tipo 14 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. O mesmo acontece com os tipos 15 e 16. Estes tipos de homógrafos podem ser considerados híbridos, no sentido em que necessitam de informações sintácticas e semânticas para a sua desambiguação.

Como se pode observar da análise da Tabela 25, outras oposições gramaticais (tipos 5, 6, 9, 11, 13) 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.

Tabela 25: Tipos de homógrafos pertencentes a classes morfossintácticas diferentes.

Tipo	Oposição gramatical e alternância vocálica	Homógrafo
1	[e] Nome / [E] Verbo	aceno, acerto, apelo, aperto, apreço, arrepelo, começo, concerto, conserto, desemprego, desespero, emprego, enredo, erro, esmero, espeto, flagelo, gelo, governo, interesse, interesses, modelo, pego, peso, pena, penas rego, remo, selo, testo, zelo
2	[o] Nome / [O] Verbo	abono, aborto, acordo, adorno, aforro, almoço, arrojo, arroto, choco, choro, conforto, consolo, contorno, controlo, 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, troço
3	[o] Nome/ [O] Verbo	rola, rolha, soma

Tabela 25: Tipos de homógrafos pertencentes a classes morfossintácticas diferentes (cont.).

4	[e] Verbo / [E] Nome	colher, meta
5	[e] Contracção / [E] Verbo	desses, deste, destes
6	[o] Verbo/ [O] Advérbio	fora
7	[e] Adj., Nome/ [E] Verbo	seco, seca, secas
8	[o] Adj., Nome/ [O] Verbo	boto[132]
9	[e] Dem. / [E] Adj., Nome	este
10	[e] Verbo / [E] Adj., Nome	leste
11	[o] Prep./ [O] Verbo	sobre
12	[@] Verbo/ [E] Nome	pegada
13	[o] Adj./ [O] Nome	rota, rotas, tola, tolas
14	[o] Nome/ [O] Nome/ [O] Verbo	corte, cortes, forma, formas, molho, soco
15	[e] Prep./ [e] Nome/ [E]Verbo	cerca
16	[e] Nome/ [E] Verbo/ Nome[E]	pega, pegas

A Tabela 26 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.

Tabela 26: Tipos de homógrafos com a mesma classe morfossintáctica.

Tipo	Oposição gramatical e alternância vocálica	Homógrafo
17	[e] Nome / [E] Nome	besta, bestas
18	[e] Nome / [E] Nome	sede, sedes
19	[e] Nome / [E] Nome	medo, medos
20	[e] Nome / [E] Nome, Verbo	termos
21	[o] Nome/ [O] Nome	cor
22	[o] Nome/ [O] Nome	lobo, lobos
23	[o] Nome/ [O] Nome	bola, bolas
24	[@] Verbo/ [E] Verbo	pregar

A oposição gramatical é essencialmente estabelecida entre Nomes com significados diferentes (excepto no tipo 24, que opõe Verbo a Verbo), ao passo que a

[132] <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*).

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 14, 15 e 16, 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.

3.3.3. Algoritmos de desambiguação

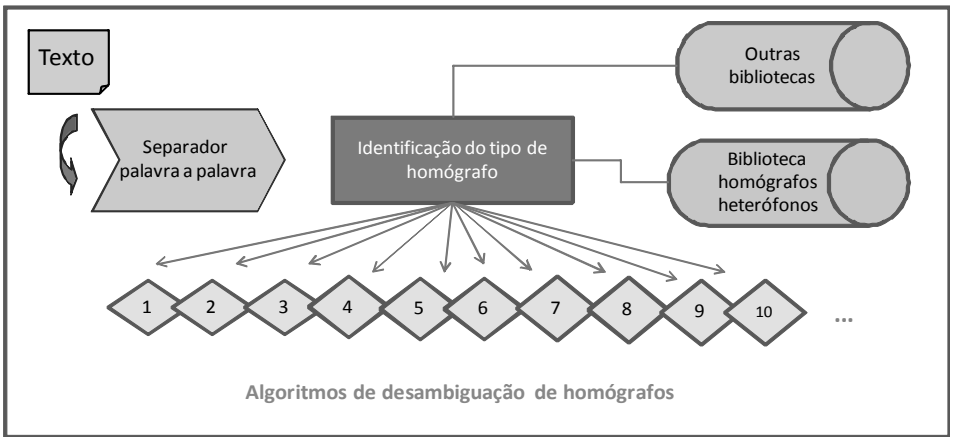

Figura 15: Funcionamento do desambiguador de homógrafos heterófonos.

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 15).

Na Tabela 27, apresentam-se os símbolos usados na representação gráfica dos algoritmos, bem como o seu significado.

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 16-40).

Tabela 27: Simbologia usada nos algoritmos.

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
PONT	, . ! ? … ; :
SP	espaço
+	seguido de
,	ou
or	condição alternativa
and	condição aditiva

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 7a e 7b (*vide* Figuras 20, 22, 23). 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)).

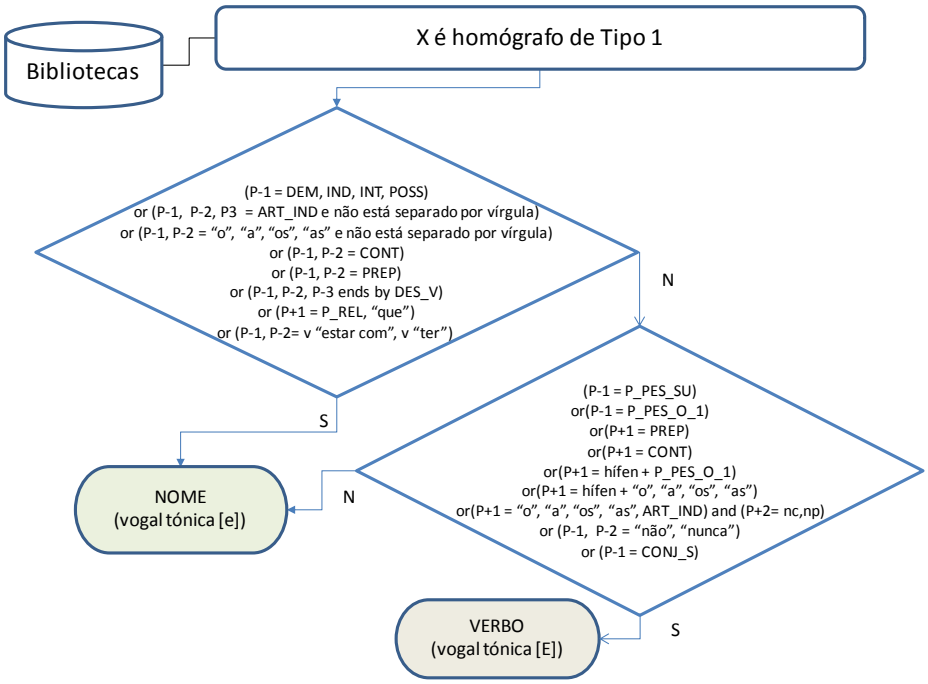

Figura 16: Algoritmo de desambiguação de homógrafos de tipo 1 (ex. <gosto>).

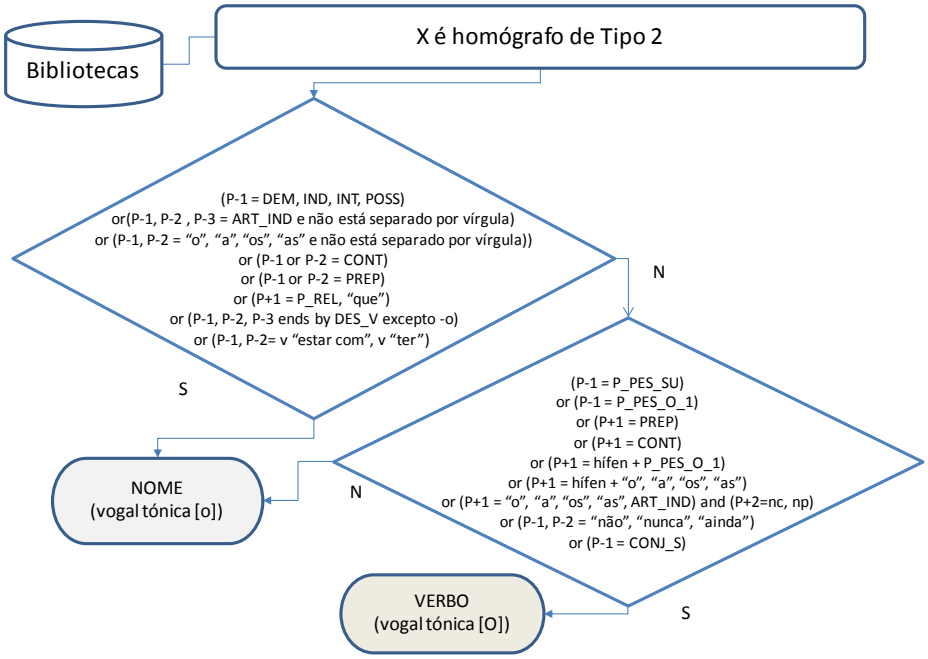

Figura 17: Algoritmo de desambiguação de homógrafos de tipo 2 (ex.<acordo>).

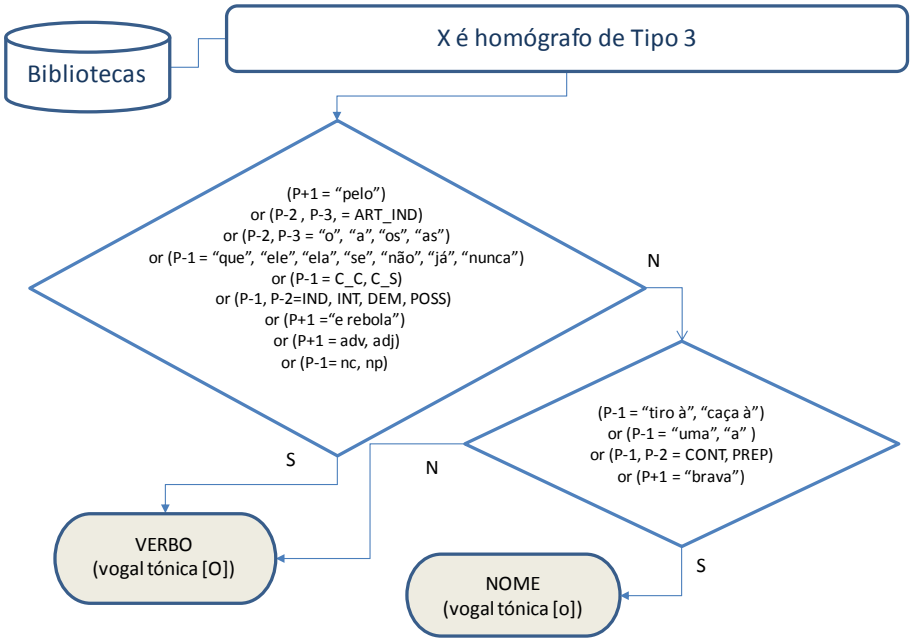

Figura 18: Algoritmo de desambiguação de homógrafos de tipo 3 (ex.<rola>).

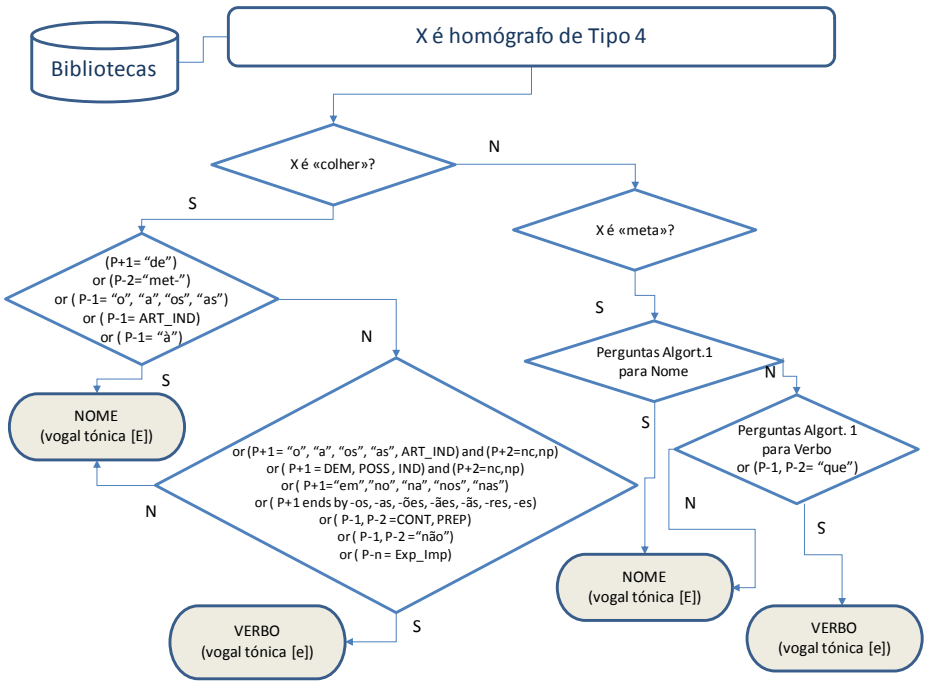

Figura 19: Algoritmo de desambiguação de homógrafos de tipo 4 (<colher>, <meta>).

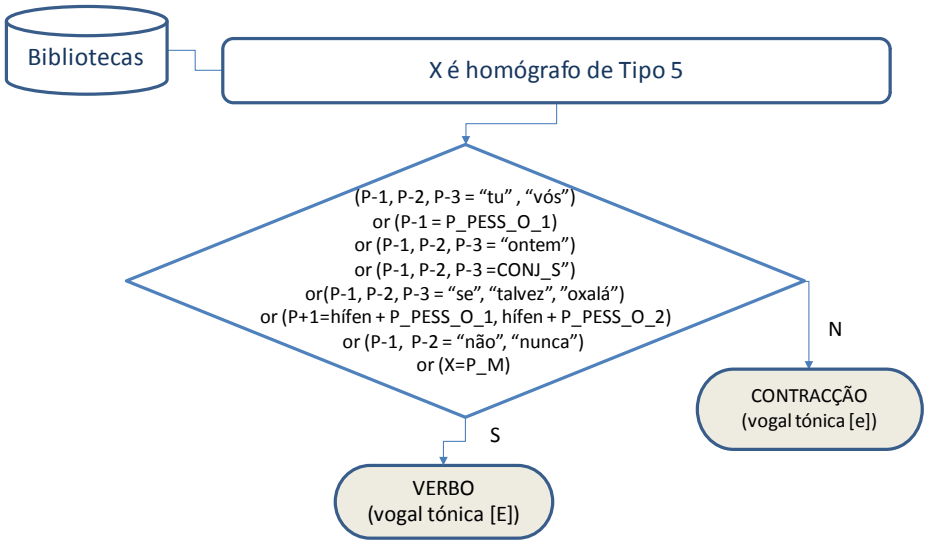

Figura 20: Algoritmo de desambiguação de homógrafos de tipo 5 (ex.<desses>).

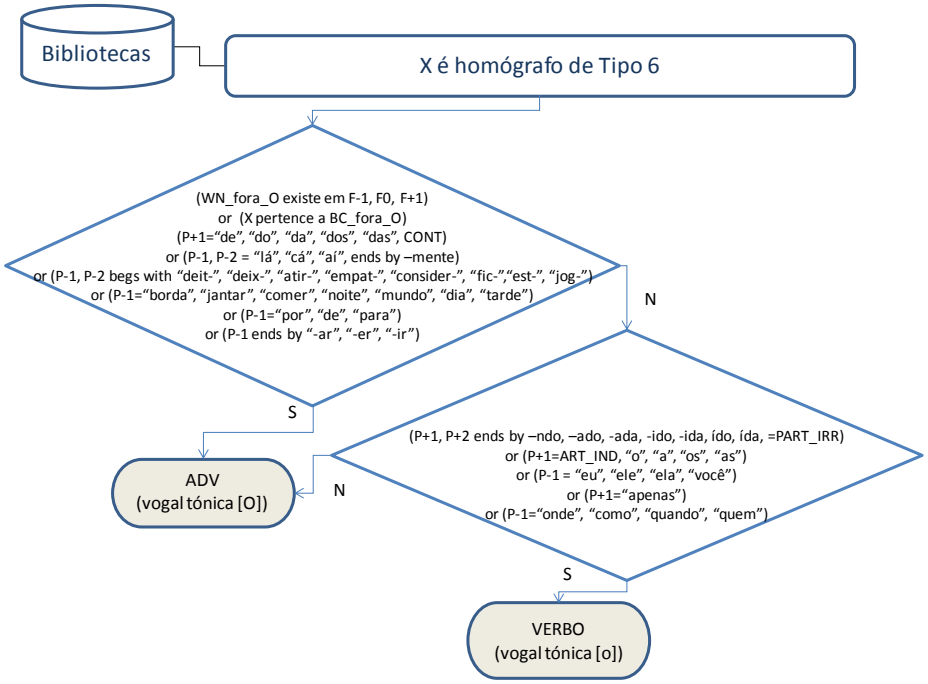

Figura 21: Algoritmo de desambiguação de homógrafos de tipo 6 (<fora>).

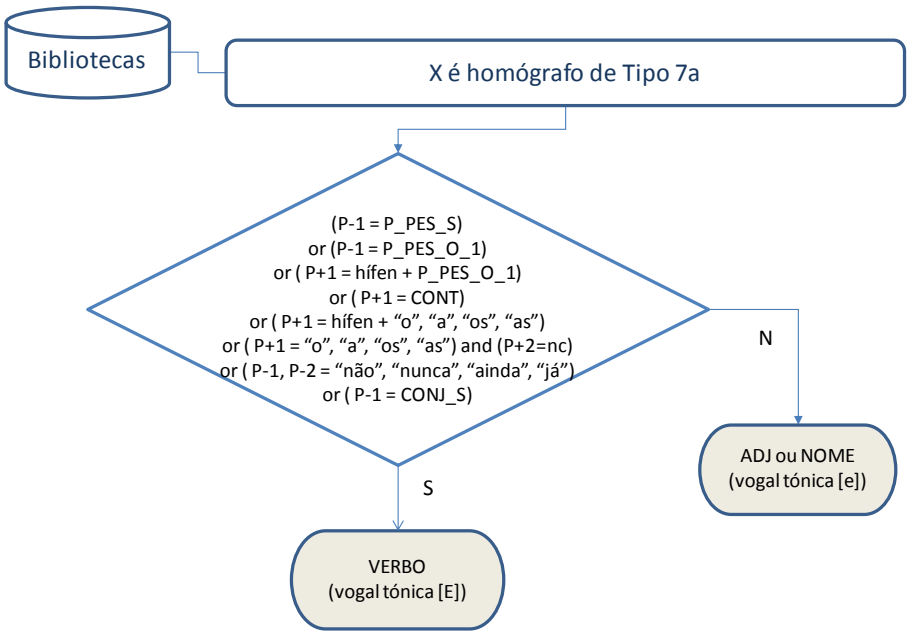

Figura 22: Algoritmo de desambiguação de homógrafos de tipo 7a (ex.<seco>).

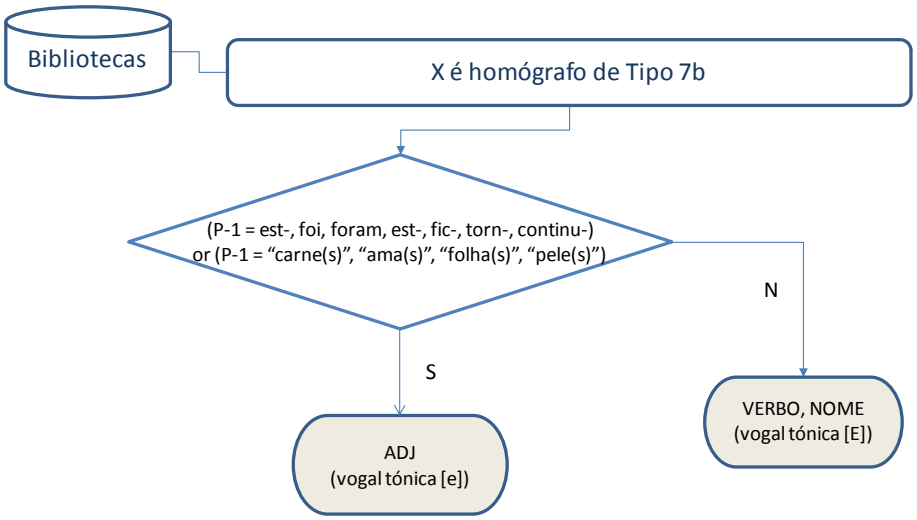

Figura 23: Algoritmo de desambiguação de homógrafos de tipo 7b (ex.<seca(s)>).

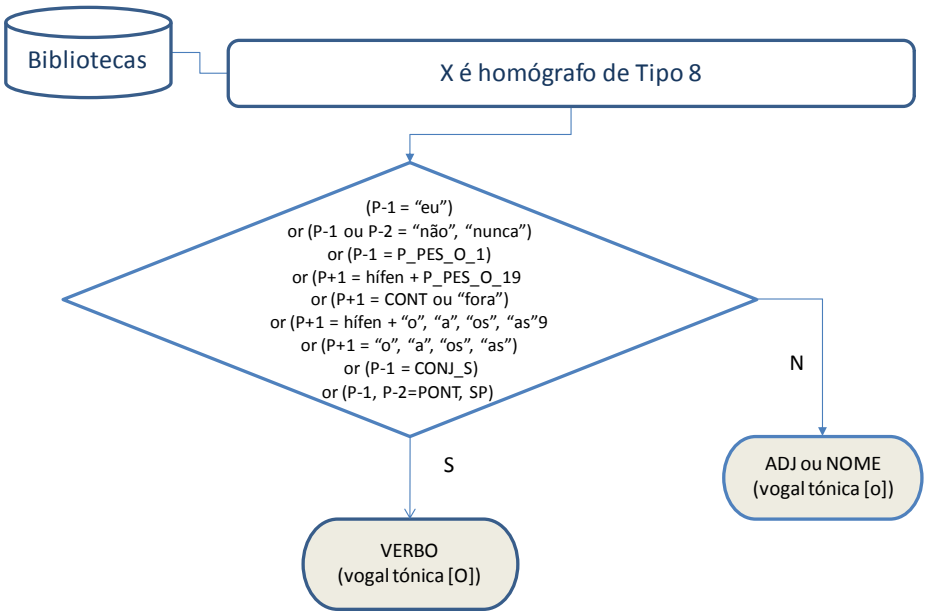

Figura 24: Algoritmo de desambiguação de homógrafos de tipo 8 (<boto>).

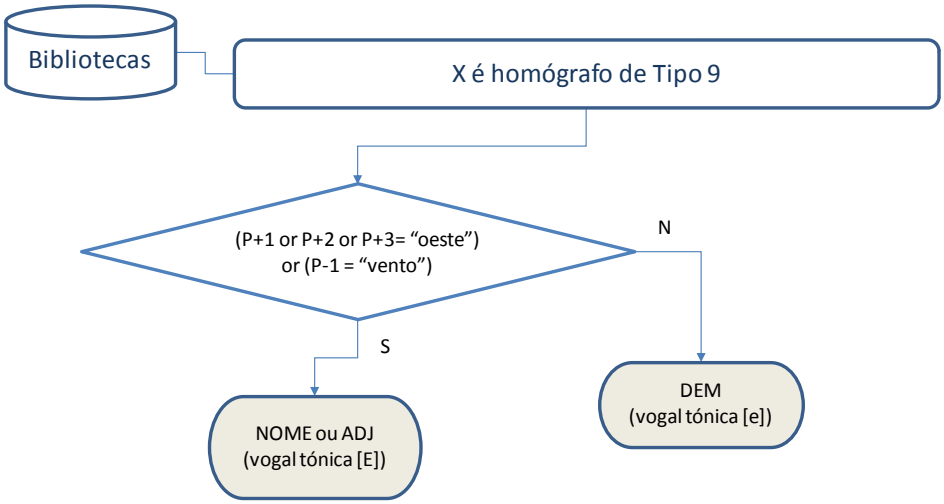

Figura 25: Algoritmo de desambiguação de homógrafos de tipo 9 (<este>).

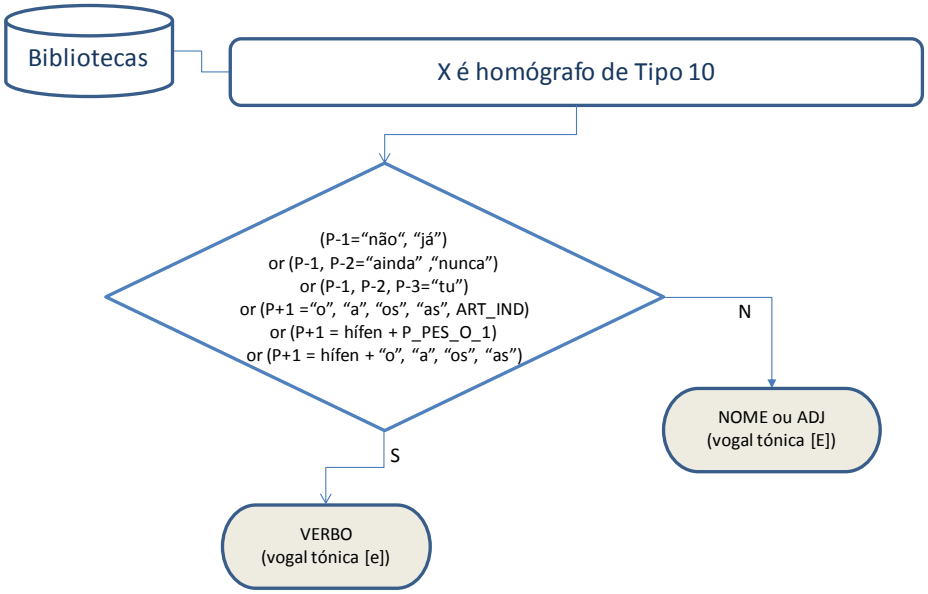

Figura 26: Algoritmo de desambiguação de homógrafos de tipo 10 (<leste>).

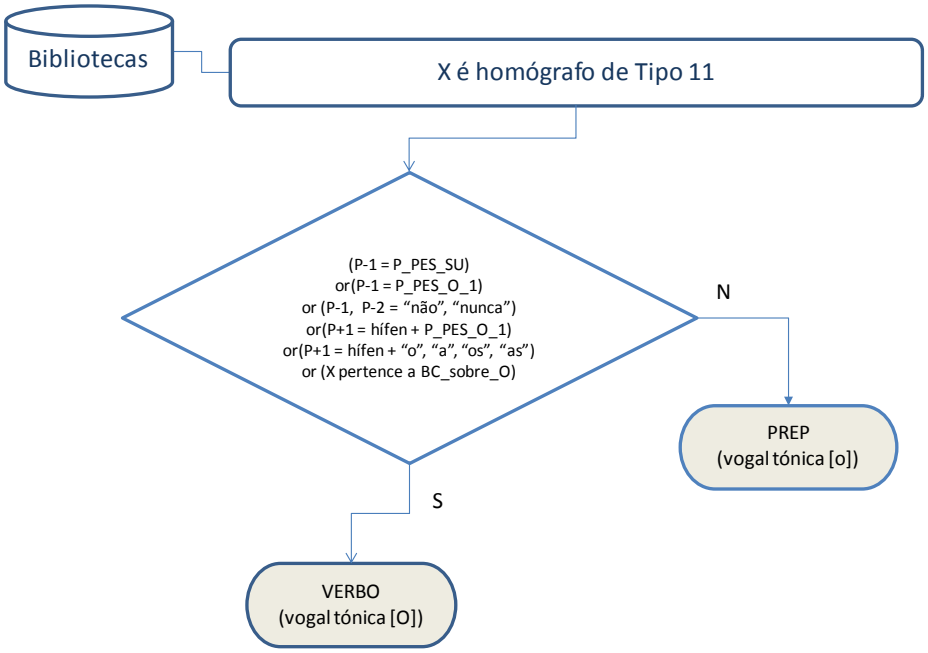

Figura 27: Algoritmo de desambiguação de homógrafos de tipo 11 (<sobre>).

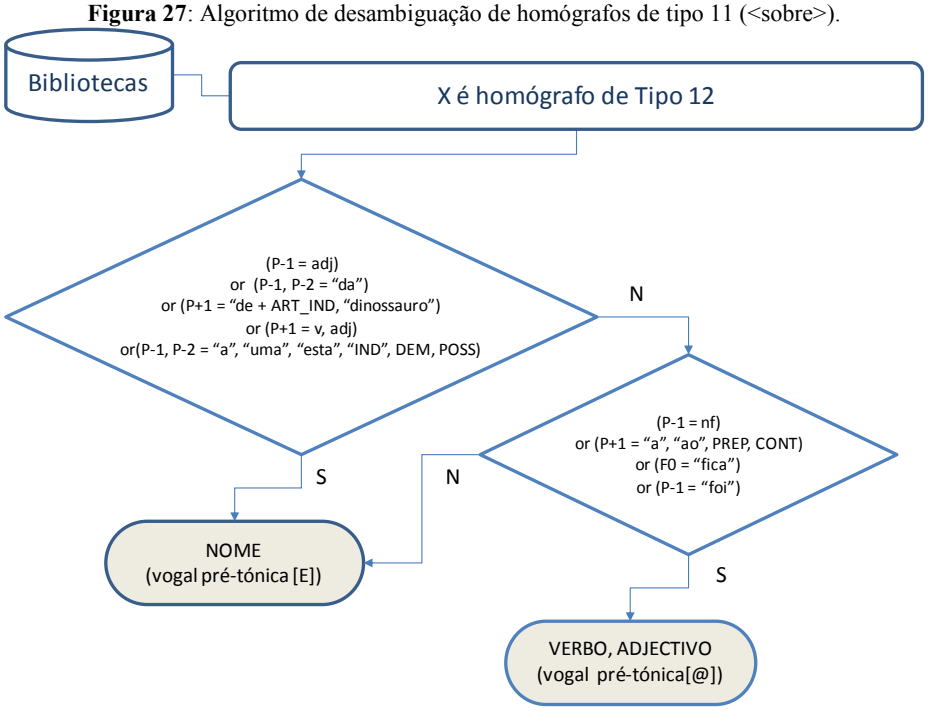

Figura 28: Algoritmo de desambiguação de homógrafos de tipo 12 (<pegada>).

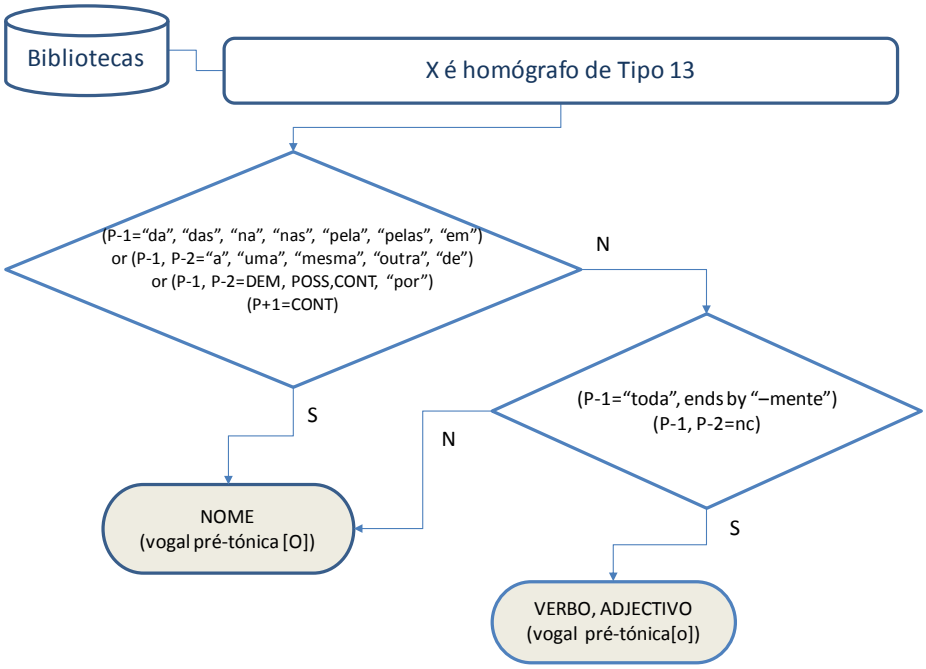

Figura 29: Algoritmo de desambiguação de homógrafos de tipo 13 (ex.<rota>).

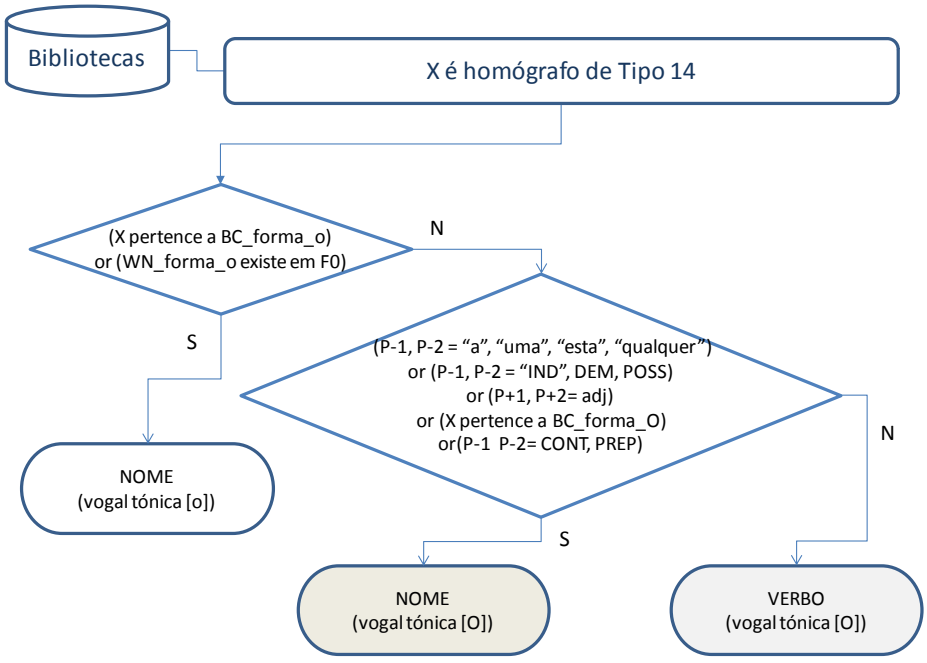

Figura 30: Algoritmo de desambiguação de homógrafos de tipo 14 (ex.<corte>).

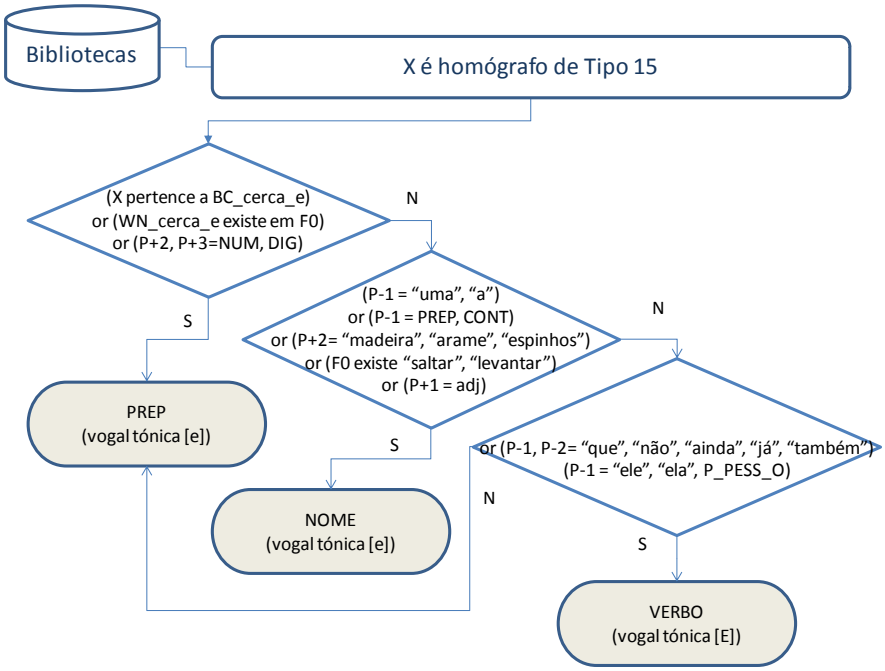

Figura 31: Algoritmo de desambiguação de homógrafos de tipo 15 (<cerca>).

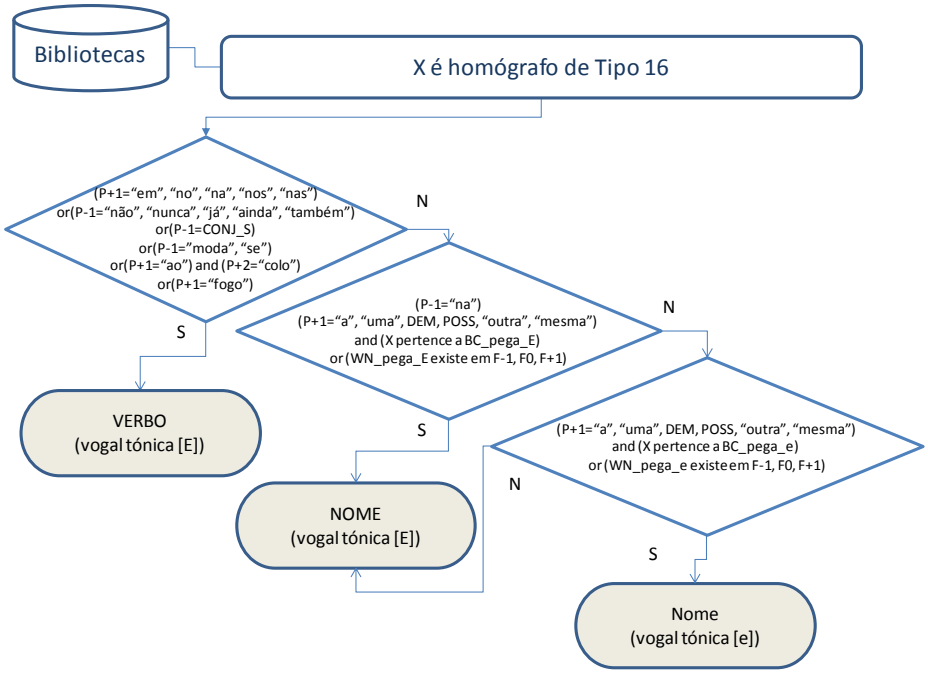

Figura 32: Algoritmo de desambiguação de homógrafos de tipo 16 (ex.<pega>).

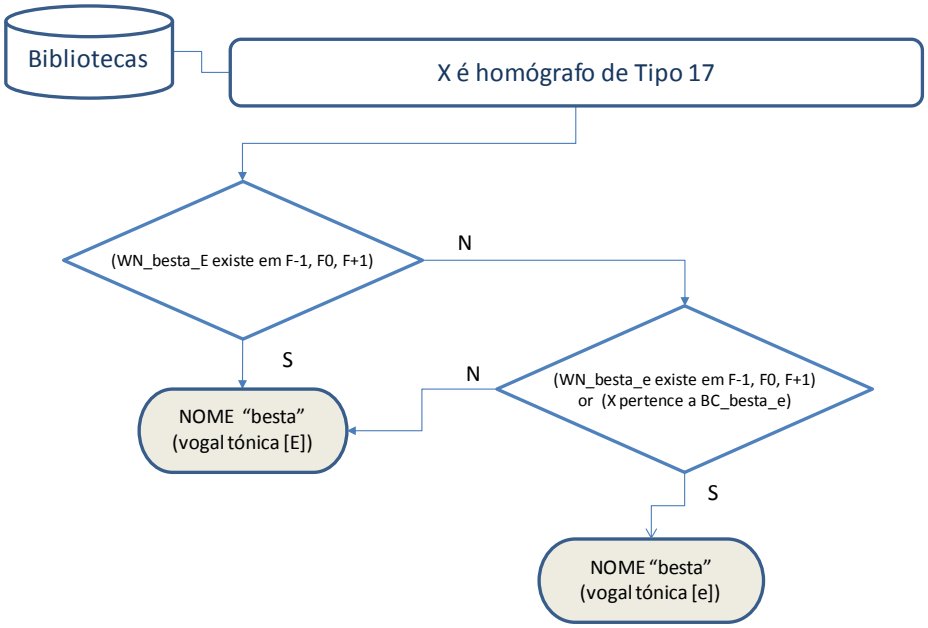

Figura 33: Algoritmo de desambiguação de homógrafos de tipo 17 (ex.<besta>).

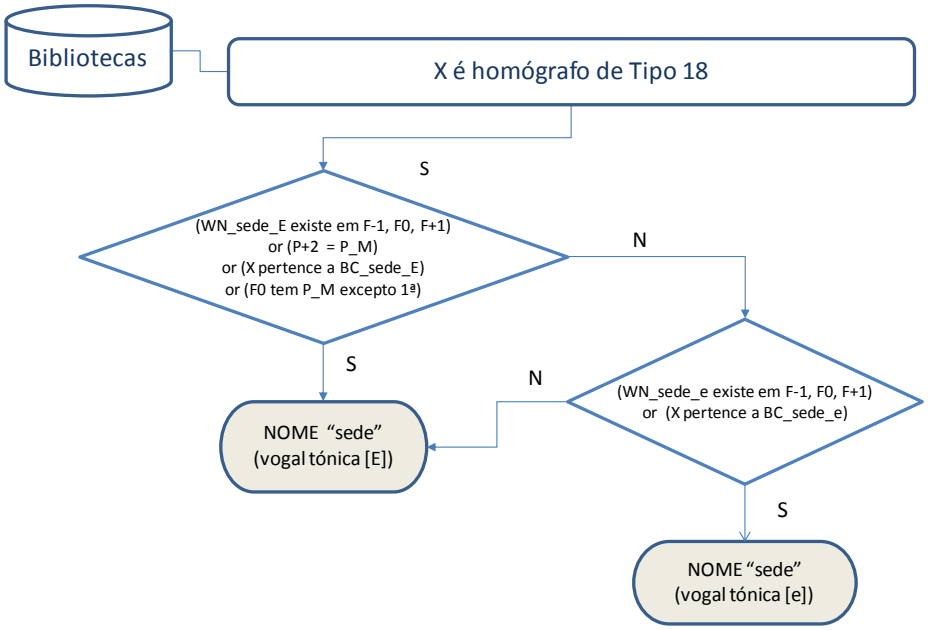

Figura 34: Algoritmo de desambiguação de homógrafos de tipo 18 (ex.<sede>).

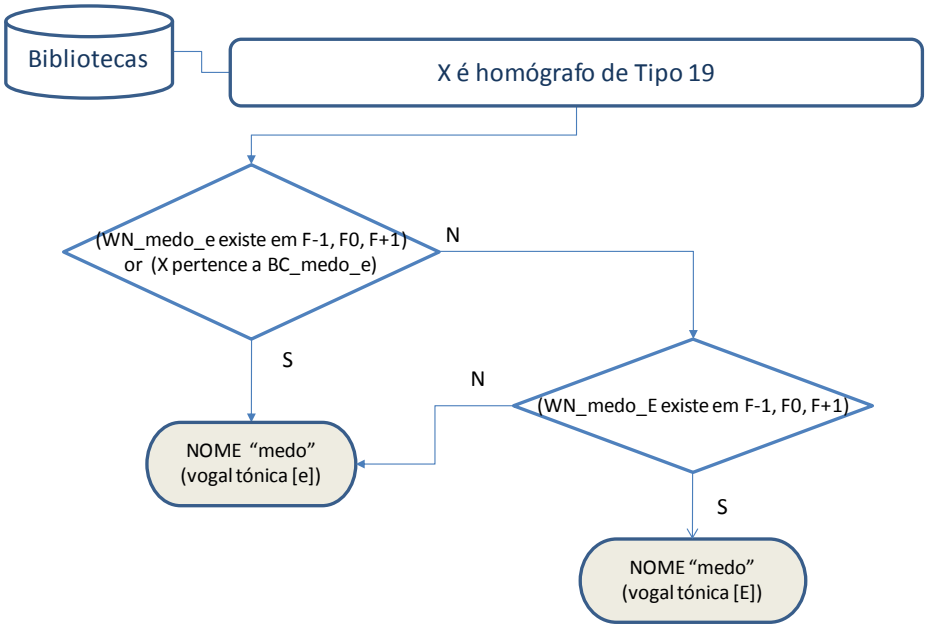

Figura 35: Algoritmo de desambiguação de homógrafos de tipo 19 (ex.<medo>).

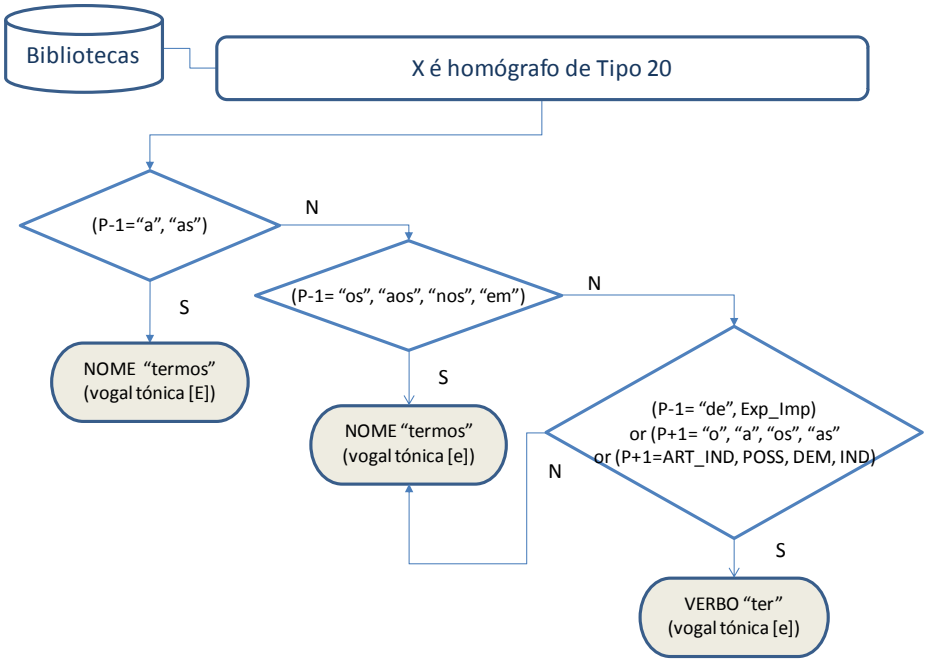

Figura 36: Algoritmo de desambiguação de homógrafos de tipo 20 (<termos>).

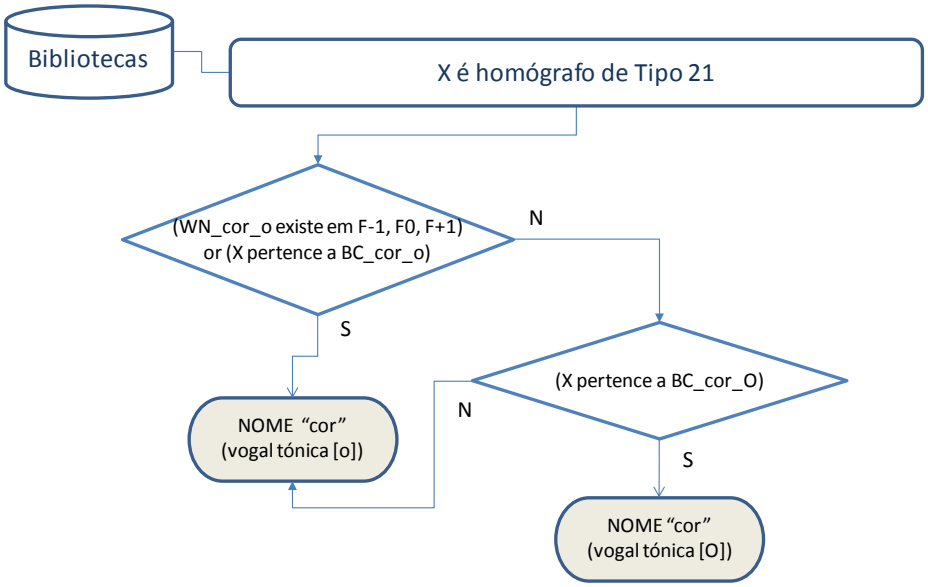

Figura 37: Algoritmo de desambiguação de homógrafos de tipo 21 (<cor>).

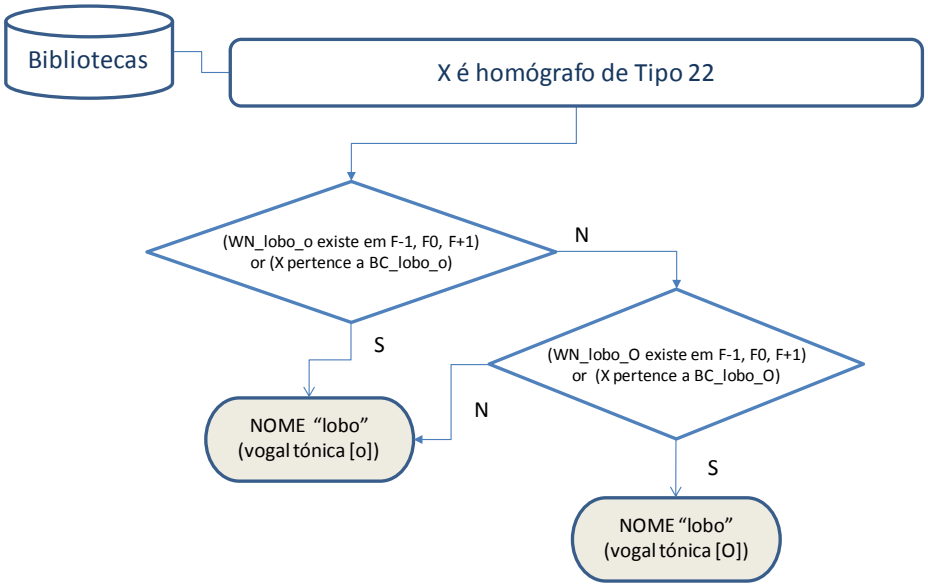

Figura 38: Algoritmo de desambiguação de homógrafos de tipo 22 (ex.<lobo>).

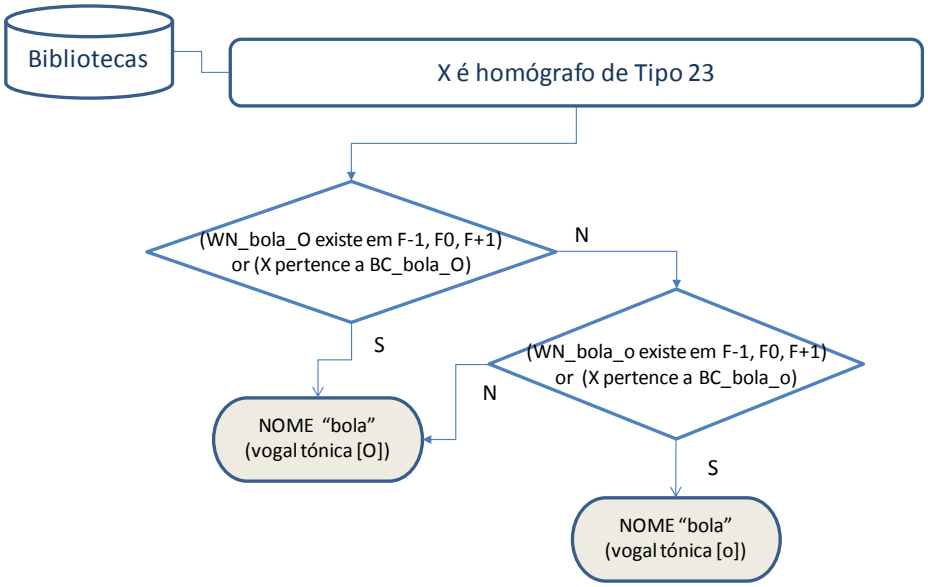

Figura 39: Algoritmo de desambiguação de homógrafos de tipo 23 (ex.<bola>).

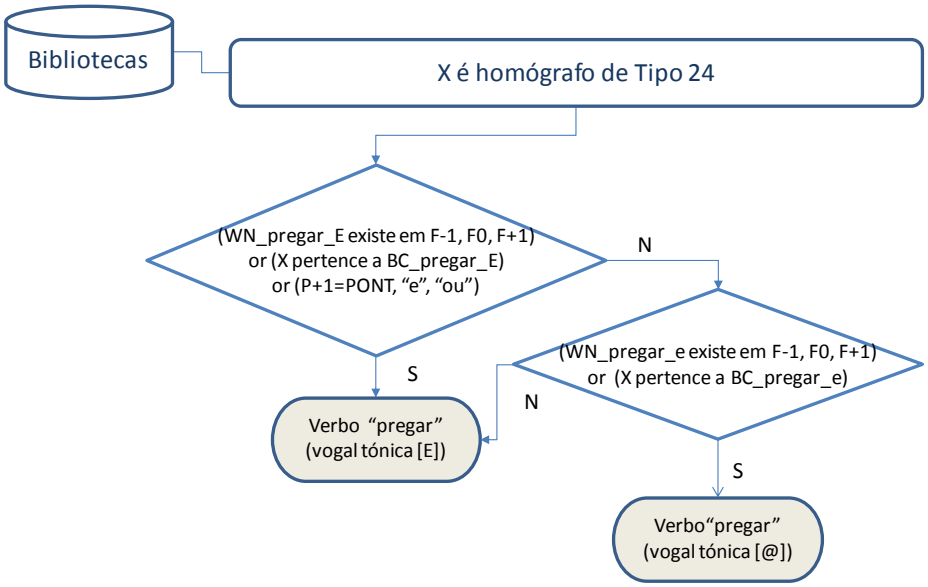

Figura 40: Algoritmo de desambiguação de homógrafos de tipo 24 (<pregar>).

Os homógrafos de tipo 7 foram entretanto separados em dois subtipos, 7a (incluindo <seco>) e 7b (para <seca> e <secas>), uma vez que se constatou que

<seco> apenas pode ser Adjectivo (<pão seco> [e]) ou Verbo (<eu seco a roupa>[E]), enquanto que <seca(s)> pode ser Adjectivo (<carne seca> [e]), Nome (<a seca assola a Somália> [E]) ou Verbo (<ela seca a roupa>[E]). Logo, os contextos sintácticos também são necessariamente diferentes, o que nos levou a reorganizar os algoritmos.

A título de exemplo, apresentamos na Tabela 28 as bibliotecas de combinatórias lexicais restritas para pregar_[@] e pregar_[E], bem como as respectivas *wordnets*.

Tabela 28: Bibliotecas de combinatórias lexicais e Wordnets de <pregar>.

BC_pregar_E	BC_pregar_@	WN_pregar_@	WN_pregar_E
pregar para outra freguesia	pregar uma partida	prego	almas
pregar um sermão	pregar partidas	roupa	benefícios
pregar aos	pregar uma estalada	etiquetas	bispo
pregar contra	pregar uma grande estalada	partida	costumes
pregar que	pregar um susto	partidas	Cristo
pregar ao vento	pregar olho		cruzada
	pregar rasteiras		democracia
	pregar uma descompostura		deserto
			Evangelho
			fé
			herege
			Igreja
			Jesus
			liberdade
			moral
			padre
			papa
			peixes
			púlpito
			Reino
			regionalização
			revolução
			rezar
			Santo António
			seminário
			unidade

3.4. 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. O desambiguador de homógrafos funciona independentemente dos demais, sendo apenas necessário correr o módulo de separação de frases e de separação de palavras previamente.

Dada a dificuldade em avaliar o desempenho de um sistema como este e em função das limitações a vários níveis dos corpora disponíveis, foram realizados 3 testes: um primeiro, com o objectivo de testar a performance dos 24 algoritmos propostos, pelo que foi escolhido um homógrafo representativo de cada tipo; um segundo, com o propósito de testar a performance do sistema com qualquer tipo de homógrafo conforme os homógrafos surgissem no *corpus*; e um terceiro, com objectivo de testar os outputs que não ocorressem no teste anterior. Para todos os testes, constituíram-se corpora específicos.

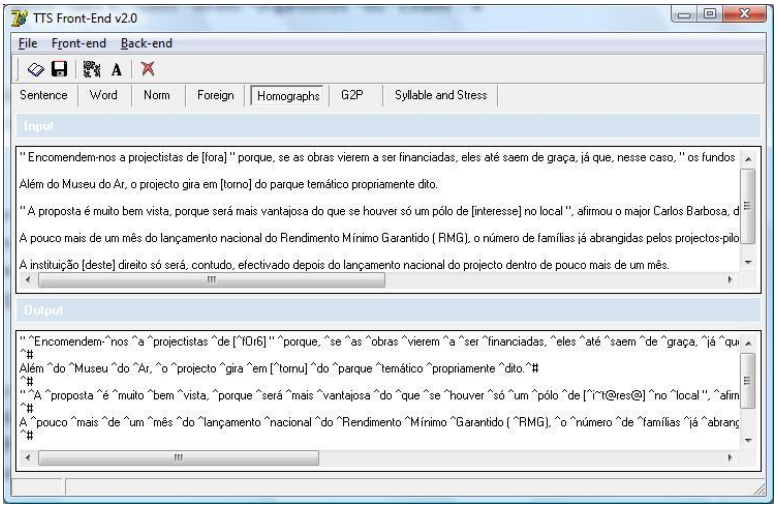

Figura 41: Interface do desambiguador de homógrafos.

Para o primeiro teste, usou-se o *corpus* Natura-Diário do Minho (Versão 3.1 do *corpus*; Versão 4.0 (21 de Fevereiro de 2006) da sua codificação)[133], que contém excertos do jornal regional Diário do Minho e que é constituído por 1738475 palavras. Seleccionou-se aleatoriamente um exemplar de cada tipo de homógrafos e fez-se uma busca no *corpus* por concordância. A busca foi realizada em 15 de Agosto de 2007. Compilou-se um documento com os resultados da busca pelo *corpus* e correu-se esse documento pelo programa. Como pode ver-se na Figura 41, que

[133] Este corpus encontra-se disponível para consulta em http://www.linguateca.pt/CETEMPublico/. Para mais informações sobre este projecto, consultar também: http://acdc.linguateca.pt/acesso/contabilizacao.html#minho (18-12-2007).

representa a interface actual do desambiguador de homógrafos, o texto de entrada foi já separado em frases (assinaladas com o carácter cardinal #) e em palavras (assinaladas com o acento circunflexo ^). Na mesma figura, pode ver-se ainda o output do desambiguador de homógrafos. Nesta fase do processamento, apenas os homógrafos são transcritos foneticamente. O output do programa foi analisado e os resultados relativos à saída fonética do desambiguador de homógrafos são apresentados na Tabela 29. O teste do sistema revelou uma taxa de erro de apenas 1,8 (ou seja, 98,2% de acerto). Este valor é bastante animador quando pensamos que a percentagem de ocorrência dos homógrafos analisados neste *corpus* é apenas de 8520 em 1738475 palavras, o que dá uma percentagem de 0,49%.

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*[134].

Tabela 29: Resultados do desambiguador de homógrafos com o teste 1.

Tipo	Homógrafo testado	# Ocorrências	# Erros	% Erros
1	'erro'	59	0	0,0
2	'gosto'	67	5	0,4
3	'rola	3	0	0,0
4	'meta'	29	0	0,0
5	'desses'	64	3	0,3
6	'fora'	primeiros 100 (em 252)	5	0,4
7	'seco'	4	1	0,1
8	'boto'	0	-	
9	'este'	primeiros 100 (em 1946)	0	0,0
10	'leste'	39	1	0,1
11	'sobre'	primeiros 100 (em 2458)	4	0,3
12	'pegada'	0	-	
13	'rota'	17	0	0,0
14	'forma'	primeiros 100 (em 1154)	0	0,0
15	'cerca'	primeiros 100 (em 1327)	0	0,0
16	'pega'	2	0	0,0
17	'besta'	0	-	
18	'sede'	primeiros 100 (em 398)	2	0,2
19	'medo'	92	0	0,0
20	'termos'	primeiros 100 (em 523)	0	0,0
21	'cor'	34	0	0,0
22	'lobo'	1	0	0,0
23	'bola'	45	0	0,0
24	'pregar'	6	0	0,0
Total	**-**	**1162**	**21**	**1,8**

[134] *Default*: saída mais frequente no total de todas as regras propostas para um dado grafema.

Outra razão tem que ver com o tipo de *corpus* usado nos testes que, pelo facto de ser jornalístico, apresenta um conjunto pouco variado de realizações de homógrafos, conduzindo quase sempre à mesma saída. Por exemplo, as saídas Verbo são raras no *corpus* Natura-Minho no que respeita aos homógrafos de tipo 1 e 2, visto que o uso da primeira pessoa verbal acarreta subjectividade, algo que não é próprio do texto jornalístico. A natureza do *corpus* explica ainda o facto de não apresentarmos resultados para os homógrafos <boto>, <pegada> e <besta>, únicos do seu tipo, simplesmente porque não ocorrem neste *corpus*.

Os erros ocorrem na desambiguação de <gosto>, <desses>, <fora>, <leste>, <sobre> e <sede>. Todos os erros encontrados decorrem do aparecimento de contextos ambíguos que caem numa condição que conduz a uma saída errada. Por exemplo, em <nós faremos gosto em vos servir>, o desambiguador devolve a saída [gOStu] porque não encontrou nenhuma das condições que levavam à saída nominal, mas encontrou a condição (P+1 = PREP), conduzindo à saída verbal.

Casos semelhantes acontecem no algoritmo de tipo 5. Em <Num caso [dEs@S], (…)>, <caso> cai na condição (P-1, P-2, P-3 =CONJ_S"), entendido como Conjunção Subordinativa e não como Nome comum, o que explica que o algoritmo o entenda como Verbo e não como Nome. Para solucionar estes casos, teria de haver um *POS tagger* mais refinado que desambiguasse previamente situações como estas. Outro caso semelhante ocorre em <se soubesse dEs@S números>, que está a cair na condição (P-1, P-2, P-3 = "se", "talvez", "oxalá"), que existe para prever orações subordinadas condicionais (iniciadas com <se>) ou orações principais com valor desiderativo (iniciadas com <oxalá>) ou hipotético (iniciadas com <talvez>), embora o output devesse ser um Nome.

Mais um caso de ambiguidade de contextos ocorreu com o homógrafo <leste> no seguinte exemplo: < Diogo de Sousa, procede a encantadora Capela dos Coimbras a qual se levanta na direcção de [leSt@] a curtos passos de distância >, em que se verificou a condição (P+1 ="o", "a", "os", "as", ART_IND), que conduzia a Verbo, embora neste caso concreto se tratasse de um Nome.

Outros erros devem-se à não verificação de nenhuma das condições previstas nos algoritmos, levando à saída *default*, que pode não ser a correcta, como no caso seguinte: <Diante dos pratos, à cabeceira, está um embrulhozinho que bem poderia confundir-se com a caixinha dos medicamentos da avó, não [fOr6] aquele laço denunciador>.

Como exemplo ainda, veja-se o erro na desambiguação de <colher>, não apresentado na Tabela 29, no seguinte exemplo: <por se destinar a [kuLEr] as domésticas desempregadas>. Este caso está a cair nas perguntas elaboradas para a saída Nome [kuLEr], em que se verifica realmente a confição (P-1= "o", "a", "os", "as"). No entanto, o algoritmo não reconhece que neste caso se trata de uma preposição e não de um artigo definido, o que provoca o erro no output.

Casos houve ainda em que a natureza dos erros tinha a ver com a ordem dos algoritmos e com o *default* que era escolhido. A performance do algoritmo de tipo 6, relativo a <fora>, melhorou substancialmente com uma alteração deste tipo (o *default* passou a ser o Advérbio e a primeira bateria de perguntas passou a ser relativa à saída Advérbio e não Verbo, como pensado inicialmente), permitindo-nos obter actualmente os valores que podem ser observados na Tabela 29.

Outros casos foram melhorados fazendo um refinamento dos algoritmos, como no caso dos homógrafos de tipo 7 <seco>, <seca> e <secas>, que foram subdivididos. Assim, <seca> e <secas> passaram a entrar no algoritmo da Figura 23, designado por 7b, mantendo-se o algoritmo 7a (Figura 22) para o homógrafo <seco>.

Antecipando críticas em relação à selecção de determinado homógrafo em detrimento de outro para avaliação do sistema, realizou-se um segundo teste, em que se extraíram as primeiras 1000 frases do *corpus* Cetem-Público contendo qualquer um dos 116 homógrafos apresentados neste capítulo[135]. Este teste permite fazer-se uma avaliação mais estendida e aleatória dos algoritmos, já que se avaliam diferentes exemplares de cada tipo. Ao mesmo tempo, este teste permite ter uma noção mais rigorosa da proporção de cada homógrafo na linguagem escrita e da sua frequência relativa em termos de outputs. O *corpus* que serviu de base ao teste 2 é assim constituído pelas primeiras 1000 frases do *corpus* Cetem-Público, divididas em 35773 palavras e 1071 homógrafos. A verificação manual do output do desambiguador de homógrafos resultou numa taxa de acerto de 97,39%, ou seja, dos 1071 homógrafos analisados, 1043 homógrafos foram correctamente transcritos. Na Tabela 30, pode ver-se o resultado deste teste: na primeira coluna lista-se por ordem alfabética os homógrafos representados no *corpus*; na segunda e terceira colunas, apresenta-se o número de realizações de cada homógrafo em função de cada output fonético; na terceira e quinta colunas estão registados os acertos efectivos do sistema para cada output. Como pode observar-se pelos dados da Tabela 30, os homógrafos mais frequentes neste *corpus* são <sobre>, <este>, <forma> e <acordo> , sendo <abono>, <arrojo>, <flagelo> ou <olho> os menos frequentes. Também se pode verificar que, apesar do grande número de ocorrências de certos homógrafos, é comum que apenas um dos outputs se verifique, o que não permite haver realizações do outro output para avaliação. Este facto tem mais uma vez que ver com a natureza jornalística do *corpus*, já que muitas realizações de homógrafos são palavras que têm maior probabilidade de ocorrer em outros tipos de texto (ex. <bola> [o]) ou em discursos na primeira pessoa, como é o caso dos homógrafos de tipo 1 e 2 com vogal tónica semi-aberta (ex. <eu choro>, <eu jogo>).

Tabela 30: Resultados do desambiguador de homógrafos com o teste 2.

homógrafo	vogal aberta	# acertos	vogal fechada	# acertos
abono			1	1
aborto			2	2
acordo			72	72
almoço			7	7
apelo			5	4
aperto			4	4
arrojo			1	1
bola	11	11		
bolas	3	3		
cerca	1	0	58	58
choro	1	0	1	1

[135] Este algoritmo foi desenvolvido por Denilson C. Silva, em trabalho não publicado, a quem eu dirijo os meus sinceros agradecimentos.

Tabela 30: Resultados do desambiguador de homógrafos com o teste 2 (continuação).

colher			1	1
começo	1	1	4	4
concerto			9	9
contorno			1	1
controlo			15	14
cor			5	5
cortes	5	5		
desemprego			9	9
desespero			3	3
desses			3	3
deste			60	55
destes			19	19
emprego			20	20
enredo			2	2
erro			5	5
esforço			3	3
este	1	0	119	119
flagelo			1	1
fora	41	41	8	8
forma	74	74		
formas	7	7		
gelo			4	4
gosto			2	2
governo			39	39
interesse	1	1	12	11
interesses			7	7
jogo			54	54
leste	4	4		
lobo			2	2
lobos			1	1
medo			10	10
meta	4	4		
modelo			9	9
olho			1	1
pena			14	14
penas			2	2
peso			9	9
piloto			9	9
pregar			2	2
reforço			11	10
rota	2	2		
rotas	1	1	1	0
secas	1	1	2	0
sede	17	16	3	3
sedes	1	1		
sobre			210	200
soma	1	1	2	2
somas			3	1
sopro			1	1
sufoco			1	1
termos			27	27
topo			7	7
torno			9	9
troço			2	2
Total	**177**	**173**	**894**	**870**

A taxa de erro resultante do teste 2 é de 2,61%, ou seja, 0,81% superior à do primeiro teste (cf. Tabela 31). Os erros mais frequentes deram-se na transcrição de <sobre>, diversas vezes transcrito como verbo, devido à ocorrência de contextos idênticos à realização de verbo. Fenómeno análogo ocorreu com a realização de <deste>, segundo maior responsável pelos erros do sistema. Estes contextos ambíguos serão tidos em consideração em futuros desenvolvimentos do nosso sistema.

Tabela 31: Resultados finais do desambiguador de homógrafos com o teste 2.

Total de homógrafos	**1071**
Total de acertos	**1043**
% acertos	**97,39**
% erros	**2,61**

Uma vez que no teste anterior nunca se verificam certas realizações de homógrafos, construiu-se um terceiro *corpus* para teste, com 5 frases representativas de cada output não encontrado no teste anterior, retiradas de diversas fontes (*internet*, livros, sugestões de pessoas que testaram o sistema em congressos) e conduziu-se um teste análogo. Os resultados, representados na Tabela 32 e 33, deram origem a 97,97% de acerto do sistema, valor que não anda muito distante dos resultados apresentados nos testes anteriores.

Tabela 32: Resultados do desambiguador de homógrafos com o teste 3.

homógrafo	**vogal aberta**	**# acertos**	**vogal fechada**	**# acertos**
abono	5	5	1	1
aborto	5	5		
acordo	5	5		
almoço	5	5		
apelo	5	5		
aperto	5	5		
arrojo	5	5		
bola			5	5
bolas			5	5
cerca	5	4		
choro	5	5		
colher	5	5		
começo	5	5		
concerto	5	5		
contorno	5	5		
controlo	5	5		
cor	5	5		
cortes	5	5		
desemprego	5	5		
desespero	5	5		
desses	5	5		
deste	5	5		
destes	5	5		
emprego	5	5		
enredo	5	5		
erro	5	5		

Tabela 32: Resultados do desambiguador de homógrafos com o teste 3 (continuação).

esforço	5	5		
este	5	5		
flagelo	5	5		
forma			5	5
formas			5	5
gelo	5	5		
gosto	5	5		
governo	5	5		
interesse	5	5		
interesses	5	5		
jogo	5	5		
lobo	5	5		
lobos	5	5		
medo	5	5		
meta			5	3
modelo	5	5		
olho	5	5		
pena	5	5		
penas	5	5		
peso	5	5		
piloto	5	5		
pregar	5	5		
reforço	5	5		
rota			5	5
rotas			5	5
secas	5	5	5	5
sedes			5	5
sobre	5	3		
somas	5	5		
sopro	5	5		
sufoco	5	5		
termos	5	4		
topo	5	5		
torno	5	5		
troço	5	5		
Total	**255**	**251**	**40**	**38**

Tabela 33: Resultados finais do desambiguador de homógrafos com o teste 3.

Total de homógrafos	**295**
Total de acertos	**289**
% acertos	**97,97**
% erros	**2,03**

Apesar de haver vários analisadores morfossintácticos aplicados ao Português (Bick, E., 2000; Ribeiro *et al*., 2003), não são conhecidas taxas de desempenho ao nível da desambiguação de homógrafos num trabalho tão amplo. Em Barbosa *et al.*, (2003d), está documentado 94,5% de acerto para o teste de apenas um par de homógrafos (<gosto>). Barbosa *et al.* (2003c), em outro trabalho, desta vez testando o homógrafo <sede>, registam 95,0% de acerto. Os nossos resultados parecem

interessantes quando comparados com os 96,45% de acerto de um desambiguador para Tailandês, em Tesprasit *et al.* (2003), usando uma técnica por "machine-learning", o Winnow, que permite considerar contextos mais ou menos alargados para a desambiguação. Seria também interessante avaliar a performance do nosso sistema ao nível do analisador morfossintáctico e semântico em próximos trabalhos.

3.5. Aplicações do sistema ao português do Brasil

Em relação à desambiguação de homógrafos em português do Brasil, e como referido atrás, destacam-se os trabalhos de Seara *et al.* (2001, 2002) e Barbosa *et al.* (2003c) e Ferrari *et al.* (2003), já descritos e comentados anteriormente.

Com o objectivo de verificar a adaptabilidade do desambiguador de homógrafos ao português do Brasil, fez-se, numa primeira fase, uma verificação da existência em PB dos 116 homógrafos heterófonos encontrados para em PE.

Tabela 34: Correspondência de homógrafos em PE e PB.

	EM PE			**EM PB**		
#	**Tipo 1**	**NOME**	**VERBO**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
1	aceno	6senu	6sEnu	Ø	N	-
2	acerto	6sertu	6sErtu	igual	S	-
3	apelo	6pelu	6pElu	igual	S	-
4	aperto	6pertu	6pErtu	igual	S	-
5	apreço	6presu	6prEsu	igual	S	-
6	arrepelo	6R@pelu	6R@pElu	Ø	N	-
7	começo	kumesu	kumEsu	igual	S	-
8	concerto	ko~sertu	ko~sErtu	igual	S	-
9	conserto	ko~sertu	ko~sErtu	igual	S	-
10	desemprego	d@ze~pregu	d@ze~prEgu	igual	S	-
11	desespero	d@z@Speru	d@z@SpEru	igual	S	-
12	emprego	e~pregu	e~prEgu	igual	S	-
13	enredo	e~Redu	e~REdu	igual	S	-
14	erro	eRu	ERu	igual	S	-
15	esmero	@Zmeru	@ZmEru	igual	S	-
16	espeto	@Spetu	@SpEtu	igual	S	-
17	flagelo	fl6Zelu	fl6ZElu	igual	S	-
18	gelo	Zelu	ZElu	igual	S	-
19	governo	guvernu	guvErnu	igual	S	-
20	interesse	i~t@res@	i~t@rEs@	igual	S	-
21	interesses	i~t@res@S	i~t@rEs@S	igual	S	-
22	modelo	mudelu	mudElu	igual	S	-
23	pena	pen6	pEn6	só ocorre Nome	N	-
24	penas	pen6S	pEn6S		N	-
25	pego	pegu	pEgu	igual	S	-
26	peso	pezu	pEzu	igual	S	-
27	rego	Regu	REgu	igual	S	-
28	remo	Remu	REmu	Ø	N	-
29	selo	selu	sElu	igual	S	-
30	testo	teStu	tEStu	igual	S	-
31	zelo	zelu	zElu	igual	S	-

Tabela 34: Correspondência de homógrafos em PE e PB (continuação).

#	Tipo 2	NOME	VERBO	Em PB	É homóg. em PB?	Alterações de grafia?
32	abono	6bonu	6bOnu	Ø	N	-
33	aborto	6bortu	6bOrtu	igual	S	-
34	acordo	6kordu	6kOrdu	igual	S	-
35	adorno	6dornu	6dOrnu	igual	S	-
36	aforro	6foRu	6fORu	igual	S	-
37	almoço	al*mosu	al*mOsu	igual	S	-
38	arrojo	6RoZu	6ROZu	igual	S	-
39	arroto	6Rotu	6ROtu	igual	S	-
40	choco	Soku	SOku	igual	S	-
41	choro	Soru	SOru	igual	S	-
42	conforto	ko~fortu	ko~fOrtu	igual	S	-
43	consolo	ko~solu	ko~sOlu	igual	S	-
44	contorno	ko~tornu	ko~tOrnu	igual	S	-
45	controlo	ko~trolu	ko~trOlu	igual	S	-
46	coro	koru	kOru	igual	S	-
47	desgosto	d@ZgoStu	d@ZgOStu	igual	S	-
48	despojo	d@SpoZu	d@SpOZu	igual	S	-
49	destroço	d@Strosu	d@StrOsu	igual	S	-
50	encosto	e~koStu	e~kOStu	igual	S	-
51	endosso	e~dosu	e~dOsu	igual	S	-
52	esforço	@Sforsu	@SfOrsu	igual	S	-
53	estorvo	@Storvu	@StOrvu	igual	S	-
54	folgo	fol*gu	fOl*gu	igual	S	-
55	gosto	goStu	gOStu	igual	S	-
56	jogo	Zogu	ZOgu	igual	S	-
57	logro	logru	lOgru	igual	S	-
58	namoro	n6moru	n6mOru	igual	S	-
59	olho	oLu	OLu	igual	S	-
60	piloto	pilotu	pilOtu	igual	S	-
61	reforço	R@forsu	R@fOrsu	igual	S	-
62	rodo	Rodu	ROdu	igual	S	-
63	rogo	Rogu	ROgu	igual	S	-
64	rolo	Rolu	ROlu	igual	S	-
65	sopro	sopru	sOpru	igual	S	-
66	suborno	subornu	subOrnu	igual	S	-
67	sufoco	sufoku	sufOku	igual	S	-
68	toco	toku	tOku	igual	S	-
69	toldo	tol*du	tOl*du	igual	S	-
70	topo	topu	tOpu	igual	S	-
71	torno	tornu	tOrnu	igual	S	-
72	troco	troku	trOku	igual	S	-
73	troço	trosu	trOsu	igual	S	-
#	Tipo 3	NOME	VERBO	Em PB	É homóg. em PB?	Alterações de grafia?
74	rola	Rol6	ROl6	igual	S	-
75	rolha	RoL6	ROL6	igual	S	-
76	soma	som6	sOm6	Ø	N	-
#	Tipo 4	VERBO	NOME	Em PB	É homóg. em PB?	Alterações de grafia?
77	colher	kuLer	kuLEr	igual	S	-
78	meta	met6	mEt6	igual	S	-

Tabela 34: Correspondência de homógrafos em PE e PB (continuação).

#	Tipo 5	CONT	VERBO	Em PB	É homóg. em PB?	Alterações de grafia?
79	desses	des@S	dEs@S	igual	S	-
80	deste	deSt@	dESt@	igual	S	-
81	destes	deSt@S	dESt@S	igual	S	-
	Tipo 6	**VERBO**	**ADV**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
82	fora	for6	fOr6	igual	S	-
#	**Tipo 7**	**ADJ, N**	**VERBO**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
83	seco	seku	sEku	igual	S	-
84	seca	sek6	sEk6	igual	S	-
85	secas	sek6S	sEk6S	igual	S	-
#	**Tipo 8**	**ADJ, N**	**VERBO**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
86	boto	botu	bOtu	igual	S	-
#	**Tipo 9**	**DEM**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
87	este	eSt@	Est@	igual	S	-
#	**Tipo 10**	**VERBO**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
88	leste	leSt@	lESt@	igual	S	-
#	**Tipo 11**	**PREP**	**VERBO**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
89	sobre	sobr@	sObr@	igual	S	-
#	**Tipo 12**	**VERBO**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
90	pegada	p@gad6	pEgad6	Ø	N	-
#	**Tipo 13**	**ADJ**	**Nome**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
91	rota	Rot6	ROt6	igual	S	-
92	rotas	Rot6S	ROt6S	igual	S	-
93	tola	tol6	tOl6	igual	S	-
94	tolas	tol6S	tOl6S	igual	S	-
#	**Tipo 14**	**NOME**	**Nome/Verbo**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
96	corte	kort@	kOrt@	igual	S	-
96	cortes	kort@S	kOrt@S	igual	S	-
97	forma	form6	fOrm6	igual	S	-
98	formas	form6S	fOrm6S	igual	S	-
99	molho	moLu	mOlu	igual	S	-
100	soco	soku	sOku	igual	S	-
#	**Tipo 15**	**PREP**	**Nome/Verbo**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
101	cerca	serk6	sErk6	igual	S	-
#	**Tipo 16**	**NOME**	**Nome/Verbo**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
102	pega	peg6	pEg6	igual	S	-
103	pegas	peg6S	pEg6S	igual	S	-

Tabela 34: Correspondência de homógrafos em PE e PB (continuação).

#	Tipo 17	NOME	NOME	Em PB	É homóg. em PB?	Alterações de grafia?
104	besta	beSt6	bESt6	igual	S	-
105	bestas	beSt6S	bESt6S	igual	S	-
#	**Tipo 18**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
106	sede	sed@	sEd@	igual	S	-
107	sedes	sed@S	sEd@S	igual	S	-
#	**Tipo 19**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
108	medo	medu	mEdu	igual	S	-
109	medos	meduS	mEduS	igual	S	-
#	**Tipo 20**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
110	termos	termuS	tErmuS	igual	S	-
#	**Tipo 21**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
111	cor	kor	kOr	igual	S	-
#	**Tipo 22**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
112	lobo	lobu	lObu	igual	S	-
113	lobos	lobuS	lObuS	igual	S	-
#	**Tipo 23**	**NOME**	**NOME**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
114	bola	bol6	bOl6	igual	S	-
115	bolas	bol6S	bOl6S	igual	S	-
#	**Tipo 24**	**VERBO**	**VERBO**	**Em PB**	**É homóg. em PB?**	**Alterações de grafia?**
116	pregar	pr@gar	prEgar	Ø	N	-

Na Tabela 34, podem ver-se os resultados da existência (S) ou não existência (N) no português do Brasil dos homógrafos encontrados em português europeu. Como se pode observar, dos 116 pares de homógrafos encontrados para o PE, 107 são igualmente presentes em PB, ou seja, verifica-se 92,24% de taxa de correspondência entre pares de homógrafos nas duas variedades. É claro que o output fonético sofre algumas adaptações sobretudo ao nível do vocalismo no PB, mas o objectivo deste exercício é apenas verificar a adaptabilidade ou não do desambiguador de homógrafos apresentado neste trabalho.

Seguidamente, o sistema foi testado sem qualquer adaptação com 500 frases extraídas do Cetem-Folha[136], contendo 11440 palavras e 525 homógrafos, usando o mesmo algoritmo desenvolvido por Denilson C. Silva, referido na secção anterior, que permite extrair todas as frases contendo qualquer um dos homógrafos listados. Foram apenas considerados na busca dessas frases apenas os 107 homógrafos existentes em PB. De um total de 525 homógrafos observados, 513 foram correctamente transcritos, o que significa que o nosso algoritmo apresenta uma taxa

[136] Disponível para consulta e download em: http://www.linguateca.pt/CETENFolha/ (19-12-2007).

de acerto de 97,71% quando testado com textos em PB, sem que seja feita qualquer adaptação (vide Tabelas 35 e 36). Esta taxa de sucesso é muito semelhante à que conseguimos para o PE, o que prova a total adaptabilidade do desambiguador de homógrafos ao PB. À semelhança do que acontecia com o PE, os maiores problemas ocorrem na transcrição de <sobre>, devido ao conflito de contextos com a forma verbal, apesar de neste tipo de *corpus* apenas se verificar a ocorrência de <sobre> como preposição, ou seja, com vogal tónica semi-fechada. Este homógrafo é também o mais frequente do *corpus*, seguido de <governo>. Outros conflitos ocorrem em <acordo> e <interesse>. É ainda curioso observar que em PB ocorrem homógrafos que não ocorrem em PE com frequência no mesmo tipo de *corpus*, como <namoro> e <selo>, o que enriquece o *corpus* de teste do nosso sistema. Como trabalho futuro, é nossa intenção estender este teste a outras realizações dos homógrafos não observadas neste *corpus*, à semelhança do que fizemos para p PE com o teste 3.

Tabela 35: Resultados do desambiguador de homógrafos com PB.

homógrafo	vogal aberta	# acertos	vogal fechada	# acertos
aborto			5	5
acordo			28	26
almoço			4	4
apelo			1	1
aperto			1	1
bola	6	6		
bolas	2	2		
cerca	1	1	23	23
choro			1	1
começo			3	3
conforto			1	1
cor			4	4
coro			1	
corte	5	5		
cortes	1	1		
desemprego			4	4
desses			5	5
deste			23	23
destes			5	5
emprego			2	2
erro			4	4
esforço			2	2
este			26	26
fora	17	17	1	1
forma	11	11		
formas	5	5		
gelo			2	2
gosto	1		1	1

Tabela 35: Resultados do desambiguador de homógrafos com PB (continuação).

governo			88	88
interesse			7	5
interesses			4	4
jogo			20	20
medo			4	4
medos			1	1
meta	2	2		
modelo			12	12
molhos	1	1		
namoro	1	1	2	2
olho			2	2
pega	1	1		
pego	1	1		
pena			5	5
penas			5	4
peso			2	2
piloto			4	4
reforço			1	1
rota	1	1		
seco			1	1
seca	1	1		
sede	5	5		
sedes	1	1		
selo			1	1
sobre			140	135
termos			10	10
torno			5	5
troco			1	1
Total	**63**	**62**	**462**	**451**

Tabela 36: Resultados finais do desambiguador de homógrafos para PB.

Total de homógrafos	**525**
Total de acertos	**513**
% acertos	**97,71**
% erros	**2,29**

3.6. Aplicações do sistema ao galego

À semelhança do que acontece com o português europeu, o tema da desambiguação de homógrafos heterófonos tem recebido pouca atenção no âmbito do processamento da fala em galego. Não se conhecem trabalhos de listagem exaustiva de homógrafos em galego nem trabalhos que prevejam a desambiguação automática de homógrafos do galego. As *Normas ortográficas e morfolóxicas do idioma galego* (2005), na secção 2.7 relativa aos casos especiais de acentuação, referem que, em palavras em que possa haver dúvidas, se usa o acento gráfico com valor de vogal aberta para distinguir palavras homógrafas, como <fóra> (adv.) e <fora> (antepret. de *ser* e *ir*). Nas mesmas "Normas", apresenta-se o inventário das palavras homógrafas em que se usa acento distintivo (cf. p.17). Mas há casos em que não há acento distintivo: "En moitos casos non é necesario facer diferenzas gráficas, porque o significado das palabras resulta claro ao apareceren en contextos diferentes: *el colle* (*o* aberto), *colle ti* (*o* pechado), *el mete* (*e* aberto), *mete ti* (*e* pechado), etc." Trata-se naturalmente de uma convenção, porque o leitor estrangeiro ou o computador não vão conseguir reconhecer os casos em que o significado das palavras e o contexto bastam para fazer a distinção vocálica do homógrafo.

Tabela 37: Homógrafos em português europeu e no galego.

	EM PE			**EM galego**		
#	**Tipo 1**	**NOME**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
1	aceno	6senu	6sEnu	2 fechados	N	-
2	acerto	6sertu	6sErtu	2 abertos	N	-
3	apelo	6pelu	6pElu	2 abertos	N	-
4	aperto	6pertu	6pErtu	2 abertos	N	-
5	apreço	6presu	6prEsu	2 fechados	N	usa-se "aprecio"
6	arrepelo	6R@pelu	6R@pElu	Ø	N	-
7	começo	kumesu	kumEsu	igual	S	grafado "comezo"
8	concerto	ko~sertu	ko~sErtu	2 abertos	N	-
9	conserto	ko~sertu	ko~sErtu	Ø	N	-
10	desemprego	d@ze~pregu	d@ze~prEgu	igual	S	-
11	desespero	d@z@Speru	d@z@SpEru	2 abertos	N	-
12	emprego	e~pregu	e~prEgu	igual	S	-
13	enredo	e~Redu	e~REdu	2 fechados	N	-
14	erro	eRu	ERu	2 abertos	N	-
15	esmero	@Zmeru	@ZmEru	2 abertos	N	-
16	espeto	@Spetu	@SpEtu	igual	S	-
17	flagelo	fl6Zelu	fl6ZElu	igual	S	flaxelo
18	gelo	Zelu	ZElu	igual	S	xelo
19	governo	guvernu	guvErnu	2 abertos	N	goberno
20	interesse	i~t@res@	i~t@rEs@	2 abertos	N	-
21	interesses	i~t@res@S	i~t@rEs@S	2 abertos	N	-
22	modelo	mudelu	mudElu	2 abertos	N	-
23	pena	pen6	pEn6	2 fechados	N	-
24	penas	pen6S	pEn6S	2 fechados	N	-
25	pego	pegu	pEgu	igual	S	-
26	peso	pezu	pEzu	2 fechados	N	-

Tabela 37: Homógrafos em português europeu e no galego (continuação).

27	rego	Regu	REgu	igual	S	-
28	remo	Remu	REmu	2 fechados	N	-
29	selo	selu	sElu	igual	S	-
30	testo	teStu	tEStu	igual	S	-
31	zelo	zelu	zElu	2 abertos	N	celo
#	**Tipo 2**	**NOME**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
32	abono	6bonu	6bOnu	2 fechados	N	-
33	aborto	6bortu	6bOrtu	igual	S	-
34	acordo	6kordu	6kOrdu	igual	S	-
35	adorno	6dornu	6dOrnu	2 fechados		-
36	aforro	6foRu	6fORu	2 fechados	N	-
37	almoço	al*mosu	al*mOsu	2 fechados	N	usa-se "almorzo"
38	arrojo	6RoZu	6ROZu	2 fechados	N	"arroxo"
39	arroto	6Rotu	6ROtu	igual	S	-
40	choco	Soku	SOku	igual	S	-
41	choro	Soru	SOru	igual	S	-
42	conforto	ko~fortu	ko~fOrtu	igual	S	-
43	consolo	ko~solu	ko~sOlu	igual	S	-
44	contorno	ko~tornu	ko~tOrnu	igual	S	-
45	controlo	ko~trolu	ko~trOlu	2 abertos	N	como nome usa-se "control"
46	coro	koru	kOru	igual	S	-
47	desgosto	d@ZgoStu	d@ZgOStu	igual	S	-
48	despojo	d@SpoZu	d@SpOZu	igual	S	despoxo
49	destroço	d@Strosu	d@StrOsu	2 fechados	N	-
50	encosto	e~koStu	e~kOStu	igual	S	destrozo
51	endosso	e~dosu	e~dOsu	igual	S	endoso
52	esforço	@Sforsu	@SfOrsu	2 fechados	N	esforzo
53	estorvo	@Storvu	@StOrvu	2 abertos	N	estorbo
54	folgo	fol*gu	fOl*gu	2 abertos	N	
55	gosto	goStu	gOStu	igual	S	também se usa "gusto"
56	jogo	Zogu	ZOgu	igual	S	xogo
57	logro	logru	lOgru	igual	S	-
58	namoro	n6moru	n6mOru	igual	S	-
59	olho	oLu	OLu	igual	S	ollo
60	piloto	pilotu	pilOtu	igual	S	-
61	reforço	R@forsu	R@fOrsu	igual	S	reforzo
62	rodo	Rodu	ROdu	igual	S	-
63	rogo	Rogu	ROgu	igual	S	-
64	rolo	Rolu	ROlu	igual	S	-
65	sopro	sopru	sOpru	2 fechados	N	-
66	suborno	subornu	subOrnu	igual	S	-
67	sufoco	sufoku	sufOku	igual	S	-
68	toco	toku	tOku	igual	S	-
69	toldo	tol*du	tOl*du	igual	S	-
70	topo	topu	tOpu	igual	S	-
71	torno	tornu	tOrnu	igual	S	-
72	troco	troku	trOku	igual	S	-
73	troço	trosu	trOsu	2 fechados	N	"trozo"
#	**Tipo 3**	**NOME**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
74	rola	Rol6	ROl6	igual	S	-

Tabela 37: Homógrafos em português europeu e no galego (continuação).

75	rolha	RoL6	ROL6	igual	S	rolla
76	soma	som6	sOm6	Ø	N	usa-se "suma"
#	**Tipo 4**	**VERBO**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
77	colher	kuLer	kuLEr	2 fechados	N	culler/coller
78	meta	met6	mEt6	2 fechados	N	
#	**Tipo 5**	**CONT**	**VERBO**	**Em galego**	**É homóg. em Gal.?**	**Alterações de grafia?**
79	desses	des@S	dEs@S	igual	S	"deses"
80	deste	deSt@	dESt@	igual	S	o V é forma dialectal
81	destes	deSt@S	dESt@S	igual	S	
#	**Tipo 6**	**VERBO**	**ADV**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
82	fora	for6	fOr6	igual	S	adv grafado "fóra"
#	**Tipo 7**	**ADJ, N**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
83	seco	seku	sEku	2 fechados	N	-
84	seca	sek6	sEk6	igual	S	-
85	secas	sek6S	sEk6S	igual	S	-
#	**Tipo 8**	**ADJ, N**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
86	boto	botu	bOtu	igual	S	-
#	**Tipo 9**	**DEM**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
87	este	eSt@	Est@	Ø	N	-
#	**Tipo 10**	**VERBO**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
88	leste	leSt@	lESt@	igual	S	verbo é forma dialectal
#	**Tipo 11**	**PREP**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
89	sobre	sobr@	sObr@	2 fechados	N	-
#	**Tipo 12**	**VERBO**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
90	pegada	p@gad6	pEgad6	igual	S	
#	**Tipo 13**	**ADJ**	**Nome**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
91	rota	Rot6	ROt6	igual	S	-
92	rotas	Rot6S	ROt6S	igual	S	-
93	tola	tol6	tOl6	igual	S	-
94	tolas	tol6S	tOl6S	igual	S	-
#	**Tipo 14**	**NOME**	**Nome/VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
96	corte	kort@	kOrt@	igual	S	-
96	cortes	kort@S	kOrt@S	igual	S	-
97	forma	form6	fOrm6	igual	S	-
98	formas	form6S	fOrm6S	igual	S	-
99	molho	moLu	mOluS	2 abertos	N	mollo

Tabela 37: Homógrafos em português europeu e no galego (continuação).

100	soco	soku	sOkuS	2 abertos	N	grafado "zoco"
#	**Tipo 15**	**PREP**	**Nome/VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
101	cerca	serk6	sErk6	igual	S	-
#	**Tipo 16**	**NOME**	**VERBO/NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
102	pega	peg6	pEg6	igual	S	-
103	pegas	peg6S	pEg6S	igual	S	-
#	**Tipo 17**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
104	besta	beSt6	bESt6	igual	S	-
105	bestas	beSt6S	bESt6S	igual	S	-
#	**Tipo 18**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
106	sede	sed@	sEd@	igual	S	-
107	sedes	sed@S	sEd@S	igual	S	-
#	**Tipo 19**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
108	medo	medu	mEdu	igual	S	-
109	medos	meduS	mEduS	igual	S	-
#	**Tipo 20**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
110	termos	termuS	tErmuS	igual	S	-
#	**Tipo 21**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
111	cor	kor	kOr	igual	S	-
#	**Tipo 22**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
112	lobo	lobu	lObu	igual	S	-
113	lobos	lobuS	lObuS	igual	S	-
#	**Tipo 23**	**NOME**	**NOME**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
114	bola	bol6	bOl6	igual	S	-
115	bolas	bol6S	bOl6S	igual	S	-
#	**Tipo 24**	**VERBO**	**VERBO**	**Em galego**	**É homóg. em gal.?**	**Alterações de grafia?**
116	pregar	pr@gar	prEgar	igual	S	-

À semelhança do que se fez para o PB, realizou-se um levantamento dos homógrafos existentes em galego, partindo do PE como referência. Na Tabela 37, podem ver-se os resultados da existência (S) ou não existência (N) no português do Brasil dos homógrafos encontrados em galego.

Da análise da Tabela 37, constata-se que a taxa de correspondência de homógrafos entre o PE e o galego é inferior à taxa de correspondência entre o PE e o PB (de 116 pares de homógrafos, 72 são iguais, ou seja, 62,06%).

Estes dados permitem concluir que a aplicação da nossa proposta ao galego pode ser seguida, sendo apenas necessárias algumas adaptações, nomeadamente ao nível das bibliotecas utilizadas (em especial das bibliotecas de classes fechadas, das bibliotecas de combinatórias lexicais restritas e *wordnets*), uma vez que as diferenças

se dão sobretudo no plano gráfico, e ao nível do inventário de pares de homógrafos, que necessita de ser expandido. Futuramente, é nossa intenção fazer esta adaptação ao galego.

3.7. Síntese do capítulo 3

Como síntese deste capítulo gostaríamos de destacar as seguintes ideias:

- A ambiguidade dos homógrafos heterófonos representa um problema de difícil resolução nos sistemas de conversão Texto-Fala sendo responsável por 0,62% de taxa de erro;
- Este módulo permite dar resposta ao problema da leitura dos homógrafos na conversão texto-fala do português porque permite efectuar desambiguação não só morfossintáctica como semântica;
- Foram propostos 24 algoritmos baseados em regras linguísticas para solucionar um elenco de 116 pares de homógrafos;
- A análise semântica, através da consulta a bibliotecas de combinatórias fixas e a bibliotecas de wordnets, aliada à análise morfossintáctica constitui a principal inovação deste módulo;
- Foram realizados 3 testes ao sistema: um primeiro, com o objectivo de testar a performance dos 24 algoritmos propostos, pelo que foi escolhido um homógrafo representativo de cada tipo; um segundo, com o propósito de testar a performance do sistema com qualquer tipo de homógrafo conforme os homógrafos surgissem no *corpus*; e um terceiro, com objectivo de testar os *outputs* que não ocorressem no teste anterior. O primeiro teste revelou 98,2% de acerto; no segundo, obteve-se 97,39% de acerto e o terceiro resultou em 97,97% de acerto;
- Apesar de haver vários analisadores morfossintácticos aplicados ao Português, não são conhecidas taxas de desempenho ao nível da desambiguação de homógrafos num trabalho tão amplo; estes resultados parecem animadores quando comparados com os 96,45% de acerto de um desambiguador para Tailandês, em Tesprasit *et al.* (2003), usando uma técnica por "machine-learning", o Winnow, que permite considerar contextos mais ou menos alargados para a desambiguação;
- Dos 116 pares de homógrafos encontrados para o PE, 107 são convergentes em PB, ou seja, verifica-se 92,24% de taxa de correspondência entre pares de homógrafos nas duas variedades;
- O sistema foi usado sem qualquer adaptação com textos reais em PB tendo-se verificado 97,71% de acerto;
- A taxa de correspondência de homógrafos entre o PE e o galego é inferior à taxa de correspondência entre o PE e o PB (de 116 pares de homógrafos, 72 são iguais, ou seja, 62,06%);
- Como trabalho futuro pretende-se adaptar este módulo ao galego.

Capítulo 4

Leitor de estrangeirismos

A leitura de estrangeirismos, a par da desambiguação de homógrafos, representa um dos problemas de mais difícil solução para a síntese da fala em português, por vários motivos, como os que a seguir passamos a elencar: 1) as diferentes origens das palavras estrangeiras, oriundas de línguas com diferentes sistemas fonológicos, tornam difícil a previsão da sua conversão fonética; 2) as palavras estrangeiras constituem uma classe aberta, em permanente expansão na língua; estão invariavelmente associadas a avanços tecnológicos e científicos e novidades de mercado, uma vez que fazemos parte de uma economia cada vez mais globalizante que nos faz chegar produtos, marcas, termos de diversos países; 3) a escassez de trabalho linguístico sobre a integração de estrangeirismos na Língua Portuguesa e a ausência de inventários actualizados e com transcrições fonéticas[137] são aspectos que não facilitam o seu tratamento computacional; 4) tal como observado por Andrade e Lavouras Lopes (2003), "a ausência de uma política nacional da língua no domínio da importação lexical" é responsável por um actual "permissivismo e ausência de reflexão teórica sobre o fenómeno dos estrangeirismos".

[137] Destacam-se três dicionários especializados de estrangeirismos para o português europeu: Costa (1990), Machado (1994) e Schmidt-Radefelt (1997), todos eles desactualizados e com palavras estrangeiras obsoletas e de uso duvidoso. A principal crítica que apresentamos a estes dicionários é a ausência de transcrição fonética das palavras listadas, o que impossibilita qualquer trabalho sistemático sobre o comportamento fonético e fonológico dos estrangeirismos em português. O projecto "Portal da Língua Portuguesa", levado a cabo pelo ILTEC (Instituto de Linguística Teórica e Computacional), contém um dicionário de estrangeirismos de fácil consulta e bastante completo. No entanto, e apesar de estar previsto haver transcrição fonética de todo o léxico disponível no Portal, segundo informação do sítio (http://www.iltec.pt/projectos/em_curso/portal.html), essa informação ainda não se encontra disponível à data de redacção deste trabalho. Para o PE, o Dicionários da língua portuguesa contemporânea da Academia das Ciências de Lisboa (Casteleiro, coord. 2001) e a 1ª edição do Grande Dicionário da Língua Portuguesa da Porto Editora (2004) apresentam transcrições fonéticas das palavras, destacando-se o primeiro também pela sua modernidade e efeito normalizador em relação ao tratamento dos estrangeirismos, como referem Andrade e Lavouras Lopes (2003). No entanto, a sua busca torna-se bastante morosa, visto se tratar de dicionários que apenas dispõem de versão em papel. Para uma revisão do tratamento dos estrangeirismos nas últimas edições do Dicionário da Língua Portuguesa da Porto Editora, veja-se Andrade e Lavouras Lopes (2003).

Neste capítulo, propomos um módulo de leitura de estrangeirismos baseado em regras linguísticas. Fez-se, numa primeira fase, um levantamento de estrangeirismos em português. Em seguida, identificaram-se as suas origens. Depois, elaboraram-se algoritmos de identificação da língua e de conversão fonética dentro do sistema fonológico da língua de origem (francês ou inglês, visto que a maior parte dos estrangeirismos provêm destas línguas), sempre tendo em conta a sua adaptação ao sistema fonológico do português, enquanto língua de chegada. O sistema foi implementado e testado, tendo-se obtido 88,05% de taxa de acerto por palavra e 98,14% de taxa de acerto por fone. Os resultados foram discutidos, bem como a sua aplicabilidade ao português do Brasil e ao galego.

Este trabalho deu origem às seguintes publicações:

- Simões, C.; Calado, A.; Braga, D.; Teixeira, C., Dias, M; "European Portuguese Accent in Non-native English models for ASR systems", *12th Iberoamerican Congress in Pattern Recognition - CIARP 2007*, Viña del Mar- Valparaíso, Chile, November 2007, pp. 738-747.
- Braga, D.; Resende Jr., Fernando Gil; Marques, M.A. 2007. "Leitor de estrangeirismos para sistemas de conversão Texto-Fala em PE", *XIII Encontro Nacional da APL*, Évora, 1-3 Outubro de 2007.

4.1. Definição do problema e estado da arte

> One could argue that, in real text, foreign words account for a small percentage of all the words, and so improvement in this area would have no significant impact on the overall accuracy of the system. However, we argue that, even if the amount of foreign names were relatively small, getting them right would substantially improve perceived synthesis quality.
> (Llitjos & Black, 2001)

Apesar de se tratar de um problema com reduzida expressão nas línguas, a correcta leitura das palavras estrangeiras ou estrangeirismos[138], aumenta substancialmente a qualidade perceptiva da síntese, tal como argumentam os autores em epígrafe.

[138] Seguimos a definição de estrangeirismo de Freitas et al. (2003): "O termo estrangeirismo aplica-se, aqui, a todas as palavras estrangeiras que não estão integradas no léxico português, de acordo com os critérios por nós definidos. Não designa, com efeito, o processo de passagem da palavra de uma língua para outra, como acontece normalmente com os termos *empréstimo* e *importação*. Por outro lado, não designa apenas a primeira fase na importação de uma lexia, como para Lavouras Lopes e Rebello d'Andrade (1997). A opção terminológica que aqui defendemos está, contudo, longe de ser original. Podemos observá-la, por exemplo, em Lavouras Lopes (1992), Rebello d'Andrade e Lavouras Lopes (2002) e também na introdução do Dicionário da língua portuguesa contemporânea da Academia das Ciências de Lisboa (Casteleiro, coord. 2001)". Deixamos ainda o conceito de estrangeirismos neste dicionário, conhecido pela sua modernidade no plano do tratamento dos estrangeirismos: "Quanto aos estrangeirismos ou neologismos externos, ou seja, vocábulos importados de línguas modernas e ainda hoje sentidos como tal, registam-se: 1) na sua forma

A leitura de palavras estrangeiras constitui outro problema de difícil resolução na área de síntese da fala em geral, e na área da síntese do português em particular, uma vez que a sua pronúncia obedece, por um lado, às regras fonológicas da língua de partida, e por outro, às regras fonológicas da língua de chegada, consoante o seu grau de integração. As diferentes origens dos estrangeirismos constituem outro problema, dado que cada língua possui o seu inventário fonológico e as suas regras de marcação de tonicidade, invariavelmente distintas das da língua de chegada. A própria identificação da palavra estrangeira representa ainda outro problema para os sistemas de conversão texto-fala, dado que muitas vezes a ortografia não basta para fazer esse reconhecimento. Finalmente, os estrangeirismos apresentam diferentes graus de integração na língua de chegada, como descrito em Freitas *et al.* (2003)[139], sendo que na segunda fase de integração, por exemplo, se verifica a possibilidade de formação de novas palavras segundo as regras morfológicas do Português, o que leva a que coexistam dois inventários fonológicos em simultâneo (ex. <surfista> [s6rfiSt6]; <checkar> [SEkar]).

Tabela 38: Estrangeirismos num *corpus* de vários tipos de texto do Expresso online.

Tipo de texto	# palavras	# estrangeirismos	% estrangeirismos
Opinião	1657	42	2,5
Economia	1278	9	0,7
Desporto	1188	16	1,3
Actualidade	1342	14	1,0
Africa	1258	19	1,5
Emprego	1170	2	0,2
Total	**7893**	**102**	**1,3**

Apesar de representarem sempre uma pequena percentagem na língua, ou seja, cerca de 1,4% do léxico[140] e 1,3% em textos (veja-se os dados da Figura 42 e da

---

de origem, os que atingiram um certo grau de generalização e aceitação, como *antidumping*, *copyright*, *design* (…); 2) na sua forma de origem, mas com remissão para a forma aportuguesada ou semi-aportuguesada proposta, por vezes já usada por alguns autores, aqueles que o uso implantou, como *abajur* (do fr. *abat-jour*), *ateliê* (do fr. *atelier*) (…); 3) na sua forma de origem, mas com remissão para um equivalente vernáculo, vocábulo ou expressão, já usual ou com possibilidade de generalização, aqueles que designam conceitos ou objectos integrantes da cultura dos nossos dias, como *avant-scène → proscénio, barbecue → churrasco, barman → empregado de bar* (…)".

[139] Segundo Freitas et al. (2003), a integração dos estrangeirismos no PE atravessa três fases passando pelos seguintes fenómenos: 1) primeira fase: adaptações fonética e morfossintáctica imediatas, monossemia: manutenção do significado com o qual a palavra é importada, grafia da língua de origem e hesitação nso tipos gráficos; 2) segunda fase: adaptações fonética e morfossintáctica progressivas, possibilidade de formação de novas palavras por composição e prefixação, formas concorrentes a nível gráfico e atestação lexicográfica, normativizada ou não; 3) terceira fase: fixação do acento fonológico, fixação do género e da forma de plural, possibilidade de derivação, polissemia com tendência para extensão, restrição ou modificação do significado da forma original e atestação lexicográfica normativizada.

[140] Num estudo realizado pela Academia das Ciências de Lisboa e referido em Casteleiro, coord. (2001, vol.I: xv), ao longo de seis anos, foram recenseados nos principais periódicos

Tabela 38), os estrangeirismos são uma classe aberta, em permanente expansão na língua, designando uma grande panóplia de entidades, desde marcas, a empresas, instituições, nomes próprios, topónimos, moedas, produtos, a termos técnicos e científicos, o que torna obrigatório o seu tratamento pelos sintetizadores de fala.

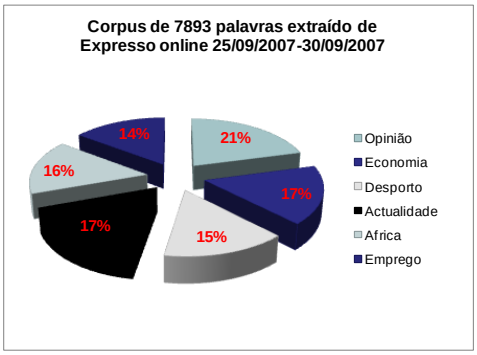

Figura 42: Percentagem dos estrangeirismos no *corpus* do Expresso online por tipo de texto.

Entre as propostas de resolução da leitura de estrangeirismos, contam-se as técnicas por dicionário (Black *et al.*, 1998), os modelos estatísticos CART-based (Llitjos & Black, 2001) e os modelos por n-grams (Chen, 2006). Muitos autores consideram o problema da leitura de estrangeirismos inserido nos seus módulos de conversão grafema-fone, havendo para este módulo uma grande variedade de técnicas disponíveis (Taylor, 2005). Poucos, contudo, se debruçam sobre este tema em especial e, quando o fazem, centram-se na leitura dos nomes próprios de origem estrangeira (Yang *et al*., 2006; Mareuil *et al.*, 2005), muito comuns em línguas que estão no cruzamento de muitas culturas, como o Inglês e o Francês.

A nível da síntese da fala em português, o assunto da leitura de estrangeirismos tem merecido pouca atenção. Em Céu Viana *et al.* (1994), fazem-se observações muito interessantes sobre o comportamento dos estrangeirismos do Português a nível fonológico e propõem-se algumas regras de conversão grafema-fone para palavras estrangeiras, com uma perspectiva ampla do problema, ou seja, independentemente da sua origem. Porém, não são apresentados resultados sobre a performance do conversor Texto-Fala ao nível dos estrangeirismos.

4.2. Leitor de estrangeirismos

Neste capítulo, apresenta-se um leitor de estrangeirismos integrado num sistema de conversão texto-fala em português europeu. A concepção deste módulo, construído

---

portuguese 4000 estrangeirismos, a maior parte na sua forma gráfica de origem, na seguinte proporção: 70% anglicismos, 20% galicismos e 10% de outras origens. Destas 4000 palavras, apenas 1000 foram inseridos no Dicionário da Academia, com cerca de 70000 entradas, o que significa que 1,42% do léxico do português são estrangeirismos.

segundo regras linguísticas, assenta na identificação da língua de origem e passa por três fases (ver Figura 43): o pré-processamento, que inclui um dicionário com transcrição fonética de palavras que não provêm do inglês ou do francês, o identificador de galicismos e anglicismos (necessário por representarem o maior número de estrangeirismos em português europeu), e dois conversores grafema-fone, um para galicismos e outro para anglicismos. Estes algoritmos apenas convertem as sequências gráficas dos estrangeirismos que apresentam transcrições fonéticas não admitidas pelo conversor grafema-fone do português, sendo as restantes sequências lidas pelo conversor grafema-fone do português.

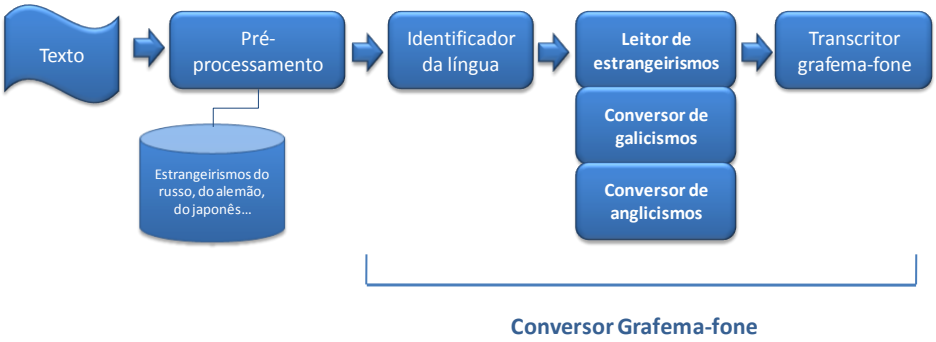

Figura 43. Arquitectura do leitor de estrangeirismos.

Uma das dificuldades iniciais deste trabalho foi a necessidade de *corpora* de análise. Para isso, elaborámos um inventário de estrangeirismos com cerca de 1000 palavras de diferentes origens a partir de dicionários especializados (Costa, 1990; Machado, 1994; Schmidt-Radefelt, 1997), dicionários electrónicos (Dicionário de Estrangeirismos do Portal da Língua Portuguesa, desenvolvido pelo ILTEC[141]), prontuários (Estrela *et al.*, 2004; Bergström & Reis, 2007), e recolhas manuais. Em seguida, separámos os estrangeirismos segundo a sua língua de origem. Foram considerados apenas os estrangeirismos que se enquadram na definição de primeira e segunda fases de integração no léxico do PE segundo a proposta de Freitas *et al.* (2003), visto que, na terceira fase, o estrangeirismo já está perfeitamente integrado aos níveis fonético, morfológico e até gráfico, sendo interpretado como uma palavra Portuguesa e seguindo directamente para o conversor grafema-fone.

Outra dificuldade, já referida atrás, foi a inexistência de reflexão teórica e de normalização no que respeita ao comportamento fonológico e fonético dos estrangeirismos em PE, o que provoca muitas dúvidas de transcrição fonética. Em caso de dúvidas e sempre que possível, consultou-se o Dicionário da Academia das Ciências de Lisboa (Casteleiro, 2001), por possuir transcrição fonética e atestado valor normalizador.

[141] Disponível em: http://www.portaldalinguaportuguesa.org/?action=estrangeirismos (21-12-2007)

Em seguida, passaremos a descrever as várias fases e funcionamento do leitor de estrangeirismos.

4.2.1. Pré-processamento das palavras estrangeiras

Nesta fase, e após a normalização do texto, o sistema vai percorrer as bibliotecas de palavras estrangeiras que não são de origem inglesa nem francesa ou que possuem uma pronúncia que escapa às regras quer do leitor de estrangeirismos, quer do conversor grafema-fone para o português europeu. Na verdade, este sub-módulo foi incluído como parte do pré-processamento ao nível da implementação. Contudo, a nível estrutural, pertence ao módulo de leitura de estrangeirismos. Constam deste módulo as palavras de origem alemã, russa, árabe ou japonesa e algumas de origem inglesa ou francesa cuja transcrição escape ao leitor de estrangeirismos. Na Tabela 39, apresentam-se as palavras estrangeiras que se encontram nesta categoria e o respectivo output fonético com marcação de sílaba tónica e divisão silábica, de forma a evitar mais processamento em outros momentos.

Tabela 39: Palavras estrangeiras que constam do pré-processamento.

Estrangeirismo	Transcrição Fonética
apartheid	a.par.ta1j.d@
ayatollah	aj6.tO1.l6
Auschwitz	a1wS.vitz
Beethoven	bE.to1.v6n
Björk	‘bjOrk
budda	bu1.d6
Dostoievski	dOS.tO1j.Evs.ki
Ericsson	E.ri1.ks6n
Ginseng	Zi1.~s6~g
Glasnost	glaS.nO1St
Güell	gwE1l*
Guggenheim	gu.g61.najm
Gulbenkian	gul*.be~1.kj6n
Ikea	i.kE1.a
Innsbruck	i.n@S.bru1k
jihad	Zi.a1d
kalashnikov	ka.la.Sni.kO1v
Kasparov	kaS.pa.rO1v
Kusturica	kuS.tu.ri1.k6
Kuweit, Kuwait	ku.a1jt
Lufthansa	luf.t6~1.z6
Mitsubishi	mi.tsu.bi1.Si
Obikwelu	O.bi.kwE1.lu
Rimbaud	R6~.bo1
Schindler	Si~1.dlEr
Schumacher	Su.ma1.k6r
Sheraton	SE1.r6.t6n
Siemens	zi1.m6n@s
Vodka	vO1.dk6
Volkswagen	vOl*ks.va1.g6n

Se o sistema encontrar alguma destas palavras no texto, devolve a transcrição fonética correspondente. Se não encontrar, passa ao módulo seguinte: o identificador da língua. Este módulo de pré-processamento pode ser expandido.

4.2.2. Identificador da língua

O objectivo do identificador de língua é classificar o candidato a estrangeirismo segundo a sua origem. Uma vez que cada língua possui o seu sistema fonológico, foram criados conversores grafema-fone (G2P) para o inglês e para o francês, tendo em conta a adaptação fonética imediata sofrida pelas palavras estrangeiras na sua primeira fase de integração no português (Freitas *et al*., 2003). Em relação ao seu funcionamento, se o sistema não identificou nenhuma palavra estrangeira que constasse da lista anterior, é accionado o identificador da língua, que começa por procurar sequências gráficas típicas de palavra estrangeira (cf. Figura 44). Se alguma das condições do primeiro losango da Figura 44 se verificar, a palavra em análise é identificada como estrangeira.

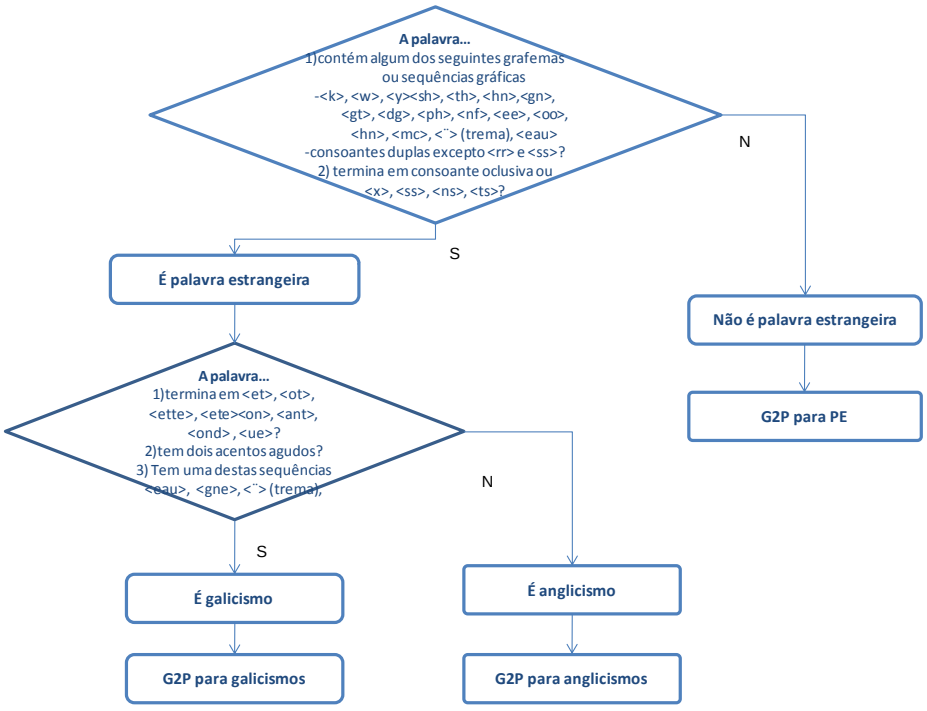

Figura 44: Algoritmo de identificação da língua.

Caso nenhuma das condições se verifique, a palavra passa para o conversor grafema-fone do PE (descrito no Capítulo 5). Se a palavra foi identificada como estrangeira, é accionada a segunda bateria de perguntas apresentada no segundo losango. Se a resposta for positiva, a palavra é identificada como galicismo, sendo

seguidamente convertida pelo conversor de galicismos. Se a resposta for negativa, a palavra é interpretada como anglicismo, sendo processada pelo conversor de anglicismos.

4.2.3. Leitor de estrangeirismos

Dado que toda a integração do estrangeirismo começa por uma adaptação fonética (Freitas *et al.*, 2003) e dado que se trata de transcrição fonética nesta fase do processamento, um primeiro exercício que se fez foi o mapeamento entre os fonemas do Inglês e os fonemas do Português[142], por um lado, e o mapeamento entre os fonemas do Francês e os fonemas do Português, por outro. Este exercício não foi simples, dado existirem pronúncias alternativas que são tanto mais próximas da língua de origem quanto o nível de proficiência em Inglês ou em Francês do falante. No entanto, enquanto uma pronúncia mais próxima da língua de origem pode ser considerada mais prestigiante de um ponto de vista sociolinguístico para palavras que ainda não se encontram integradas no léxico, o mesmo não acontece para palavras da terceira fase de integração, podendo até ter conotações negativas:

> A pronúncia mais próxima da língua de origem é considerada mais prestigiante pelo facto de poder evidenciar um grau de cultura ou conhecimento mais elevado. No entanto, esse tipo de conservadorismo poderá implicar conotações sociolinguísticas negativas, caso se verifique em relação aos fenómenos característicos da primeira fase de integração ou ocorra em relação a palavras da terceira fase, palavras já integradas no léxico. (Freitas *et al.*, 2003)

A partir da segunda fase de integração, começam a aparecer grafias aportuguesadas em coexistência com as grafias de origem. Neste caso, as palavras não são interpretadas como estrangeirismos, passando para o G2P do PE. No entanto, há casos de coexistência dos dois sistemas fonológicos (ex. <icebergue>), que constam da tabela de excepções (Tabela 40).

Tabela 40: Tabela de anglicismos e galicismos que constituem excepção.

Estrangeirismo	Transcrição Fonética
airbus	Er.b61s
au ralenti	O.Ra.le~.ti1
bacon	be1j.k6n
badminton, badmington	bad.mi~1.t6n
barman	ba1r.mEn
beatles	bi1.t6l*s
blaser, blazer	ble1j.z6r
boxe	bO1.ks@
boxer	bO1.ks6r

[142] Este trabalho esteve também na base do artigo sobre a pronúncia da Língua Inglesa falada por falantes portugueses, na área do Reconhecimento de Voz: Simões et al. 2007.

Tabela 40: Tabela de anglicismos e galicismos que constituem excepção (continuação).

budget	b61.dZ6t
bus	b61s
business	biz.n61s
cameraman	ka1.m@.r6.mE1n
car	ka1r
cd	se.de1
center	se~1.t6r
charter	Sa1r.t6r
deadline	dE.d@.la1j.n@
dealer	di1.l6r
décor	dE1.kOr
delete	di.li1.t@
disc drive	di1Sk.draj.v@
e-mail	i.m61jl*
excel	EksE1l*
fan	f6~1
features	fi.tS61.r@S
ferryboat, ferry-boat	fE1.Ri.bo1w.t@
file	fa1j.l@
game	ge1j.m@
gangster	g6~.g@S.tE1r
gentleman	Ze~.t@l*.mE1n
gin	Zi~1
ginger-ale	Zi~.Z6.r61j.l@
gospel	gO1S.p6l*
hamburguer	6~bu1rg6r
hard-core	ar.d@.kO1.r@
Hertz	E1rtz
hi-fi	aj.fa1j
hotmail	O.t@.m61jl*
ice tea	aj.s@.ti1
icebergue[143]	aj.s@.bE1r.g@
inter-rail	i~.tEr.R61jl*
laser	l61j.zEr
leasing	li1.zi~g
made in	me1j.din
mail	m61jl*
mails	m61j.l@S
manager	mE1.n6.Z6r
mass media	mas.mE1.dj6
modem	mO1.dE.m@
mouse	ma1w.z@

[143] Nesta palavra coexistem os sistemas fonológicos do inglês em <ice> e do português em <bergue>.

Tabela 40: Tabela de anglicismos e galicismos que constituem excepção (continuação).

movie	mu1.vi
nonsense, non-sense	nO.n@.se~1.s@
online	O.n@.la1j.n@
outsider, out-sider	au.t@.sa1j.dEr
palace	pa1.l6.s@
party	pa1r.ti
pen	pE1.n@
performance	pEr.fO1r.m6~.s@
portable device	pO1r.t6.b6l*.di.va1j.s@
raid	Ra1jd
rail	Ra1jl*
rent-a-car, rent a car	Re~.t6.ka1r
roaming	Ro1w.mi~g
roll on	RO.lO1.n@
rouge	Ru1.Z@
router	Ru1.t6r
save	s61j.v@
self-service	sE1l*f.s61r.vi.s@
sex-appeal	sE.ksa.pi1l*
Shakespeare	Sej.k@S.pi1.6r
slide	sla1jd@
soul	so1wl*
spider man	spaj.d6r.m61n
star	sta1r
state of the art, state-of-the-art	st61j.t@.Of.di.a1rt
stereo	stE1.rju
Stones	sto1w.n@S
striper	stri1.p6r
strip-tease	strip.ti1.z@
superman, super man	su.p6r.m61n
surf	S61rf
teenager	ti.ne1j.Z6r
tête-à-tête	tE1.t@.a.tE1.t@
time	ta1j.m@
t-shirt	ti.S61rt
tour	tu1r
under-score	6~.d6r.scO1.r@
wireless	waj.6r.lE1s

Ainda em Freitas *et al.* (2003), descrevem-se os principais fenómenos de adaptação fonética imediata dos anglicismos no português, que sintetizamos em seguida:

- Consoantes nasais em posição pré-consonântica são associados às vogais precedentes, nasalizando-as (ex. fra[n]chising → fr[6~]chising);
- Neutralização da distinção fonológica entre vogais breves e longas (ex. d[i:]ler → d[i]ler);
- Simplificação das africadas [tS] e [dZ] (ex. chat → [S]at; jeans → [Z]eans);

- Substituição da aproximante central alveolar inglesa pela vibrante alveolar dentro da palavra (ex. t-shirt → t-shi[r]t) ou pela vibrante uvular em início de palavra (ranking → [R]anking).

A partir da análise do nosso inventário e das opções fonéticas sugeridas pelo Dicionário da Academia das Ciências de Lisboa (Casteleiro, 2001), podemos acrescentar os seguintes fenómenos de adaptação fonética imediata ao nível dos anglicismos, para além dos apresentados antes:

- Supressão da fricativa faringal inglesa, i.e., do <h> aspirado em início de palavra (ex: hip hop → [Ø]ip[Ø]op);
- Substituição da fricativa interdental inglesa pela oclusiva dental surda (ex. thriller → [t]riller) ou sonora (ex. Big Brother → Big Bro[d]er) ou pela fricativa dental surda (ex. Bluetooth → Bluetoo[s]);
- Simplificação das consoantes duplas[144] (ex. coffee-break → co[f]ee-break);
- Paragoge de um schwa [@] no final de palavras terminadas por grafemas que não ocorrem em Português, como <p>, <t>, <k>, <b>, <d>, <g>, <f>, <h>, <x> (ex. <clip>, <budget>, <punk>, <band>, <blog>, <bluff>, <crash>, <relax>)[145];
- Vocalização ou semivocalização consoante os contextos de <y> (ex. baby-doll → bab[i]-doll; array → arra[j]);
- Total adaptação do vocalismo tónico e átono do Inglês (ex. surf → s[6]rf; brownie → br[aw]n[i]; cheesecake → ch[i]sec[6j]ke)

Após a identificação da palavra como estrangeira, o sistema começa por transcrever foneticamente as consoantes, partindo de uma regra prévia: todas as consoantes duplas se simplificam, excepto <rr> e <ss>. Em seguida, aplicam-se as regras da Tabela 41. As regras começam pelas sequências gráficas mais raras, terminando com um *default*.

Tabela 41: Tabela de conversão das consoantes inglesas e francesas.

#	padrão gráfico de <b>	fone	exemplo
1	... <b>...	[b]	symbol, snob, boutique
#	**padrão gráfico de <c>**	**fone**	**exemplo**
1	…<c k>...	[k]	stock, cockpit
2	... < c > < e, i >...	[s]	center, deficit
3	…<c h>…	[S]	chat, chalet
4	<cc>	[ks]	Access
5	…<c>…	[k]	connect, disc jokey, cabaret, cognac

[144] Este fenómeno já tinha sido mencionado em Céu Viana et al. (1994).
[145] Este fenómeno também está atestado em Céu Viana et al. (1994).

Tabela 41: Tabela de conversão das consoantes inglesas e francesas (continuação).

#	padrão gráfico de <d>	fone	exemplo
1	…<d>…	[d]	deficit, hard drive
#	**padrão gráfico de <f>**	**fone**	**exemplo**
1	... < f >...	[f]	offline, free-lancer
#	**padrão gráfico de <g>**	**fone**	**exemplo**
1	... < g > < e, é, i >...	[Z]	Exchange, rouge
2	... < g u > < e, i >...	[g]	hambúrguer
3	…<g n>…	[J]	champagne, champignon
4	…<ng> <SP, Pont, s>…	[~g]	ping pong
5	... < g >...	[g]	Groove, engagé
#	**padrão gráfico de <h>**	**fone**	**exemplo**
1	... < h >...	[ ]	hip hop
#	**padrão gráfico de <j>**	**fone**	**exemplo**
1	... < j >...	[Z]	disc jokey
#	**padrão gráfico de <k>**	**fone**	**Exemplo**
1	... <k>...	[k]	Ketchup
#	**padrão gráfico de <l>**	**fone**	**Exemplo**
1	... < l > <C/h, Pont>...	[l*]	holding, hall, gospel
2	...<l>...	[l]	lifting
#	**padrão gráfico de <m>**	**fone**	**Exemplo**
1	…<mc>…	[mEk]	Mc Donalds
2	... < m >...	[m]	mailbox, modem
#	**padrão gráfico de <n>**	**fone**	**Exemplo**
1	... <n>...	[n]	nonstop,walkman
#	**padrão gráfico de <p>**	**fone**	**Exemplo**
1	...<ph>...	[f]	Geographic, photo
2	... < p >...	[p]	ping-pong
#	**padrão gráfico de <q>**	**fone**	**Exemplo**
1	... < q u >< i, e> ...	[kw][146]	Queens, Quick Silver
2	... < q >...	[k]	quark
#	**padrão gráfico de <r>**	**fone**	**Exemplo**
1	... < r r > ...	[R]	Ferry
2	...<(W_bgn) r>...	[R]	Rock
3	... < r >...	[r]	cross, Broadway

[146] Excepção: <quilovolt> [ki.lO.vO1l*t], quilowatt> [ki.lO.wO1t].

Tabela 41: Tabela de conversão das consoantes inglesas e francesas (continuação).

#	padrão gráfico de <s>	fone	Exemplo
1	... <s h>...	[S]	off-shore
2	...<(W_bgn) s>...	[s]	scanner, suite
3	... <V> <s> <V>...	[z]	close-up, vison
4	…<ss>…	[s]	croissant, stress, acess
5	…<s><C_UV[147]>…	[S]	casting, desktop
6	…<s><SP, Pont>…	[Ø]	tabela de excepções de palavras francesas[148]
7	…<V><s><SP, Pont>…	[S]	Optimus, corn flakes
8	…<C><s><SP, Pont>…	[s]	Barclays, Philips
9	…<s>…	[s]	outsider
#	**padrão gráfico de <t>**	**fone**	**Exemplo**
1	…<V>< th><V>…	[d]	Big brother
2	…<t h> <Pont, SP>	[s]	bluetooth
3	…<g h t><Pont, SP>…	[t]	Copyright
4	…<t h>…	[t]	aparthotel, thriler
5	…<t>…	[t]	cockpit, tour
#	**padrão gráfico de <v>**	**fone**	**exemplo**
1	... < v >...	[v]	drive-in, souvenir
#	**padrão gráfico de <w>**	**fone**	**exemplo**
1	... <w>...	[w][149]	windows, workshop
#	**padrão gráfico de <x>**	**fone**	**exemplo**
1	…<x>…	[ks]	outbox, sexy
#	**padrão gráfico de <y>**	**fone**	**exemplo**
1	…<(W_bgn)y><V>…	[j]	yankee
2	…<V><y><Pont, SP, s>	[j]	airways, array
3	... <y> <C> ...	[i]	Yguaçu
4	... <y> ...	[i]	brandy, body, baby-doll
#	**padrão gráfico de <z>**	**fone**	**exemplo**
1	... < z >...	[z]	zapping, blazer

Após a conversão das consoantes, o sistema passa à conversão grafema-fone das vogais. A estrutura das regras para as vogais é análoga à das consoantes, começando pelas sequências gráficas mais raras e terminando na saída *default*. Na Tabela 42, podem ver-se as regras de conversão grafema-fone para os anglicismos.

[147] Grafemas consonânticos não vozeados ou surdos: <p>,<t>, <k>, <q>, <c>.
[148] Lista de palavras cujo <s> final não se lê: <ménage à trois> [me.na.Za.trwa1], <collants> [kO.l6~1S], <croissants> [krwa.s6~1S].
[149] Em palavras de origem germânica, <w> articula-se [v]: <wagner, wagneriano, wálchia>.

Tabela 42: Tabela de conversão das vogais inglesas.

#	padrão gráfico de <a>	fone	exemplo
1	… < a i>	[E]	airbag, airways, fairplay
2	…<a><l l> <SP, Pont, s>…	[O]	hall, conf call
3	…<(W_bgn) C><C><a><C><V>…	[6j]	shave, braveheart
4	…<(W_bgn) C><a><C><V, y>…	[6j]	baby, bacon, take-away
5	… <(W_bgn) C><a><ck, nd, sh, rr>…	[E]	back-up, band, flash, cash & carry
6	…<are><SP, Pont, p, s>…	[Er@]	hardware, tupperware, sharepoint
7	…<an><SP, Pont, s>…	[6n]	autopullman
8	…<a n><C>…	[6~]	franchising, yang, stand
9	…<a y>…	[6j]	spray, take-away
10	…<ae><SP, Pont>…	[6j]	sundae, reggae
11	…<(W_bgn) w a>…	[wO]	walkman, Washington
12	…<a>…	[a]	fax, zapping, squash
	padrão gráfico de <e>	**fone**	**exemplo**
1	…<e e>..	[i]	cheesecake, coffee-break, feedback, feeling, jeep
2	…<b><r><e a><k>…	[ej]	breakdance, coffee-break
	…<e (m,n)><C>…	[e~]	send
3	…<e a>…	[E][150]	overhead, sweater[151]
4	…<C><e><r, t> <SP, Pont, s>…	[6]	browser, babysitter, big brother, serial-killer, gadget
5	…<e> <l> <SP, Pont, s>…	[E]	cocker spaniel, gospel
6	…<e><C><C>…	[E]	Express, best seller
7	<e><C, SP, Pont >	[@]	pickles, puzzle, cheesecake
8	…<e>…	[@]	Braveheart, cameramen
#	**padrão gráfico de <i>**	**Fone**	**exemplo**
1	…<(W_bgn) i><C\n>…	[aj]	ice tea, iceberg, i-pod
2	…<i><ne, me, ght><SP, Pont, s>…	[aj]	full-time, very-light
3	…<i n>…	[i~]	intranet, antidoping
4	…<i><n><SP, Pont, s>…	[i]	skin-head
5	…<i e> <SP, Pont, s>…	[i]	brownie, hippie
6	…<i>…	[i]	bit, flip-flop, kick-off
#	**padrão gráfico de <o>**	**Fone**	**exemplo**
1	…<oo>…	[u]	boomerang, bluetooth, zoom
2	…<o u n><C>…	[a~w~]	country, background
3	…<o u><C>…	[aw]	check-out, cowboy
4	…<o w><C>…	[aw]	brownie[152]
5	…<o a>…	[O][153]	bodyboard, broadway

[150] Excepção: em <leasing> ,<sex-appeal>, <Shakespeare>, <features>, <strip-tease><ea> lê-se [i].
[151] Excepção: em <breaveheart>, <ea> lê-se [a].
[152] Excepção: <show> [So1w], <snowboard> [snow.bO1rd].
[153] Excepção: lê-se [ow] em <ferryboat>, <roaming>.

Tabela 42: Tabela de conversão das vogais inglesas (continuação).

6	…<o><ck>…	[O]	cocker, cocktail
7	…<(W_bgn) o><C/f>…	[ow]	open, old-fashion
8	…<o><f><f>…	[O]	office, off-line
9	…<o n><SP, Pont, s>…	[On]	Scorpions, Simpsons
10	…<o><r><SP, Pont, s>…	[6]	communicator
11	…<o>…	[O][154]	golf, hip hop
#	**padrão gráfico de <u>**	**fone**	**exemplo**
1	…<(W_bgn) C><u><C>…	[6][155]	bus, bluff, budget, blush, puzzle, surf
2	…<C/q,g><u e>…	[u]	bluetooth
3	…<u><p>…	[6]	setup, ketchup, tupperware
4	…<u> <ll>…	[u]	autopullman, full-time
5	…<u>…	[u]	hamburguer

Em relação aos galicismos, cada vez em menor número no português, destacam-se os seguintes fenómenos de adaptação fonética imediata:

- Elevação das vogais nasais, visto que todas as vogais nasais do Português são [- baixas] (ex. chaper[O~] → chaper[o~])
- Vibrante uvular é realizada como dental (ex. c[R]oquis → c[r]oquis)
- Total adaptação do vocalismo tónico e átono do Francês (ex. affaire→ aff[E]r[@]; buffet → b[u]ff[e])
- Simplificação das consoantes duplas (ex. collants → co[l]ants)

Na Tabela 43, apresentamos as regras de conversão grafema-fone para as vogais dos galicismos.

Tabela 43: Tabela de conversão das vogais francesas.

#	padrão gráfico de <a>	fone	exemplo
1	….<ant><SP, Pont, s>…	[6~]	avant-lettre
2	…<au>…	[o]	au-ralenti, chauffage
3	…<ai>…	[E]	affaire
4	…<a, à>…	[a]	affaire
#	**padrão gráfico de <e>**	**fone**	**exemplo**
1	…<é, ê>…	[E]	tête-à-tête
2	…<e n><C>…	[6~]	au ralenti, engagé
3	…<e><tte><Pont, SP, s>…	[E]	manette
4	…<e (t,s)> <SP, Pont, s>…	[e]	cabaret, gourmet, guichet
5	…<(e,é) s> <SP, Pont>…	[e]	Elyseés, negligé
6	…<e><SP, Pont>…	[@]	toilette
7	…<e>…	[E]	negligé

[154] Excepção: lê-se [6] em <motherboard> [m6.d6r.bO1rd].
[155] Excepção: lê-se [u] em <duty free> [du1ti.fri1], <juke-box> [ju1.k@.bO.ks@].

Tabela 43: Tabela de conversão das vogais francesas (continuação).

#	padrão gráfico de <i>	fone	exemplo
1	….<i, ï>…	[i]	naïf, tricot
#	**padrão gráfico de <o>**	**fone**	**exemplo**
1	…<o n>…	[o~]	napperon
2	….<o t> <SP, Pont, s>…	[o]	Camelot, tarot
3	…<o i> <C>…	[wa]	Soirée, toilette
4	…<o u><C>…	[u]	boutique
5	…<o>…	[O]	cocotte, cognac
#	**padrão gráfico de <u>**	**fone**	**exemplo**
1	…<u>…	[u]	suite

Dado que a grafia de muitas palavras do inglês e do francês é de base etimológica e não fonológica e tendo-se verificado que existem palavras cujas transcrições escapam a estas regras, na Tabela 40 apresentam-se os estrangeirismos que constituem excepção e sua respectiva transcrição fonética. Encontram-se nessa tabela também os estrangeirismos que não são identificados pelo identificador da língua por não conterem as sequências fonéticas que são perguntadas pelo sistema.

No leitor de estrangeirismos, a maior dificuldade foi a marcação do acento tónico e a divisão silábica da palavra estrangeira. Segundo confirmam Freitas *et al.* (2003), a fixação do acento fonológico é uma das mudanças ocorridas ao nível do estrangeirismo na sua integração no léxico do português. E, naturalmente, essa fixação estabelece-se segundo as regras de marcação do acento fonológico do português. Uma vez que em português as palavras são paroxítonas, espera-se um comportamento semelhante em relação aos estrangeirismos. Este comportamento junta-se à tendência para acrescentar um schwa em palavras terminadas por consoante diferente de <s>, <r>, <l>, <m> ou <n>, acrescentando-lhes mais uma sílaba e frequentemente tornando-as paroxítonas. A Tabela 44 ilustra precisamente estas transformações fonológicas:

Tabela 44: Adaptação do acento fonológico e da estrutura silábica na 2ª fase de integração do anglicismo.

Acento no Inglês	Acento no Português e divisão silábica
áirbag	[Er-b´E-g@]
déadline	[dE-d@-´laj-n@]
ínternet	[i~-tEr-´nE-t@]
sóftware	[sOf-tu-´E- r@]
wórkshop	[wor-k@-S´O-p@]

A utilização directa do divisor silábico e do marcador de tonicidade sobre os estrangeirismos revelou-se pouco viável, dada a elevada taxa de erro: em 100 estrangeirismos escolhidos aleatoriamente do nosso *corpus*, o sistema errou 15% na separação silábica e 28% na marcação de sílaba tónica. É necessário um estudo mais profundo sobre a estrutura fonológica dos estrangeirismos e sua adaptação fonética ao português de forma a permitir um afinamento desses algoritmos às palavras estrangeiras.

4.3. Testes e discussão de resultados

Os vários módulos do leitor de estrangeirismos apresentados foram implementados e testados. Tendo por objectivo avaliar a importância deste módulo na arquitectura global do *front-end* do sintetizador, corremos um conjunto de 586 palavras estrangeiras 3773 caracteres sem espaços no transcritor grafema-fone isoladamente. Este corpus, diferente do que foi usado para formular os algoritmos, é composto por uma lista de palavras estrangeiras consideradas mais frequentes e incluídas num léxico fonético, essencialmente nomes de produtos, marcas, vocabulário comum, fornecido por uma empresa de software. A taxa de erro deste módulo por palavra foi de 75,4%.

Em seguida, corremos as mesmas 586 palavras no leitor de estrangeirismos proposto neste capítulo, tendo a taxa de erro descido para 11,95%. Se se considerarem os erros ao nível do fone, a taxa de erro reduz-se para 1,86%. A Tabela 45 ilustra os resultados deste teste.

Tabela 45: Resultados da avaliação do leitor de estrangeirismos.

Tipo de erro	#	%WER[156]	%PER[157]
<a>	12	2,05	0,32
<e>	22	3,75	0,58
<i>	8	1,37	0,21
<o>	15	2,56	0,40
<u>	3	0,51	0,08
<s>	3	0,51	0,08
<t>	3	0,51	0,08
<n>	2	0,34	0,05
<y>	2	0,34	0,05
Total	**70**	**11,95**	**1,86**

Os principais erros ocorrem na transcrição das vogais inglesas <e> (ex. <greatest hits> [grEtESt its], <sky news> [ski nEwS]), <a> (ex. <jackpot>[Zakpo]) e <o> (ex. <outdoor> [awtdur]). Em 1,37% dos casos (correspondentes a 8 erros repartidos pelas vogais <e> e <o> seguidas de <t> em posição final de palavra), ocorreu uma confusão de anglicismos com galicismos, em palavras com contextos gráficos comuns (ex. <internet>, <briget>, <fox-trot>, <jackpot>). Outros erros ocorrem na transcrição de consoantes <s> (ex. <Microsoft> [mikrOzOft]), <t> (ex. <national> [nEtional*]), <n> (ex. <scanner> [sk6~6r]) e <y> (ex. <sky news> [ski nEwS]).

[156] WER – word error rate (taxa de erro por palavra).
[157] PER – Phone error rate (taxa de erro por fone).

4.4. Aplicações do sistema ao português do Brasil e ao galego

Se o tema dos estrangeirismos desperta pouco interesse no que se refere ao PE, ainda menos literatura especializada se encontra sobre o processo de incorporação de estrangeirismos em PB ou em galego. Para além da falta de dicionários que descrevam as realizações fonéticas dos estrangeirismos em qualquer uma das línguas[158], também não existem normas claras quanto ao seu uso, grafia e até pronúncia.

Os grandes dicionários Houaiss (Houaiss, 2001) e Aurélio (Ferreira *et al.*, 2004) contêm transcrição fonética de estrangeirismos, mas trata-se de uma transcrição muito próxima da língua de origem. Esta opção é, em nosso entender, um pouco afastada do uso da língua, na medida em que não só não descreve a adaptação fonética ao Português do Brasil que os estrangeirismos sofrem, como descreve apenas a articulação de um grupo muito restrito de brasileiros que falam línguas estrangeiras sem sotaque. Ainda em relação ao PB, em *1001 estrangeirismos de uso corrente em nosso cotidiano* (Nicola *et al.*, 2003), encontramos estrangeirismos com informação de pronúncia, mas registada de forma ortográfica, o que torna a informação pouco rigorosa.

No que respeita ao Galego, os dicionários Xerais da Lingua Galega (Carballeira, 2000) e da Real Academia Galega (García & González, 1997)[159] incluem alguns estrangeirismos.

Na verdade, a globalização trouxe uma entrada muito rápida de estrangeirismos no Português e no Galego, o que explica a inexistência de regulamentação oficial sobre o assunto.

No entanto, o debate em torno deste assunto é aceso no Brasil e na Galiza, mais do que em Portugal, no sentido da "defesa" da língua, do purismo, no sentido da substituição do estrangeirismo por uma palavra ou expressão na língua de chegada, como se pode verificar nos excertos que se seguem:

> Em 2002 foi realizado o Seminário Agronegócio de Exportação. O Itamaraty, patrocinador do evento, exigiu o uso de agronegócio em vez de agrobusiness, que era o termo preferido pelos empresários do setor. Ponto para o Itamaraty. Sem querer ser purista, devemos defender a Língua Portuguesa. Não há necessidade alguma de usarmos palavras como startar ou mesmo estartar. Por que não iniciar, começar ou principiar? Outra palavra muito em moda é paper. Além de mal traduzido, ainda está sendo usado num sentido muito amplo. Tudo virou paper. Quando me pedem um paper, nunca sei se é um relatório, um fax, uma carta ou uma proposta. Só falta o paper higiênico. (…) para qualquer novo estrangeirismo, primeiro devemos buscar uma palavra correspondente em português. E antes de usarmos a forma estrangeira, ainda devemos tentar o aportuguesamento. (Sérgio Nogueira, "O uso das palavras estrangeiras", in http://www.portugues.com.br/art2.htm, 20-12-2007)

---

[158] A Academia Brasileira de Letras dispõe de um serviço de busca de palavras estrangeiras, em http://www.academia.org.br/abl/cgi/cgilua.exe/sys/start.htm?sid=24 (20-12-2007), porém sem qualquer informação da sua pronúncia.

[159] Disponível online em: http://www.edu.xunta.es/diccionarios/BuscaTermo.jsp (20-12-2007).

> No estudio do léxico é onde mais se pode observar a deterioración ou a vitalidade dunha língua. Aquela língua que saiba adaptarse ós novos tempos, que sexa áxil na creación de novas palavras que nomeen novos obxectos, será unha lingua com futuro. Pola contra, aquela que dependa excessivamente doutra, da que recolla todo o léxico que apareza para obxectos novos ou para os novos avances será unha língua estancada, dependente e en decadência.
> O galego non só tem que supera-lo atranco da introducción frecuente de anglicismos actualmente, senón que ademais tem que superar, ainda, unha tradición castelanizante de séculos. (Tarrío Barreiro & Seoane García, 1997)

São distintas, porém, as realidades linguísticas no Brasil e na Galiza, como aliás os excertos acima transcritos deixam perceber. Enquanto o Brasil está fortemente exposto à influência da economia americana, o que se traduz numa grande permeabilidade à entrada de termos do inglês, não apenas relacionados com as novas tecnologias, o galego debate-se antes de mais contra séculos de castelhanização, com efeitos a todos os níveis linguísticos, não apenas no léxico[160]. Actualmente, e à semelhança do que acontece com o português, a maior parte dos estrangeirismos do galego, exceptuando os castelhanismos, provém do inglês (Freixeiro Mato, 2006: 315)[161].

O que encontramos na geografia da lusofonia é uma ausência de política linguística mais assertiva em relação ao uso e pronúncia dos estrangeirismos, o que provoca nos falantes muitas dúvidas e variações ortográficas. A prova disto mesmo está na flutuação gráfica de palavras estrangeiras. Em PB, coexistem as formas <stress> e <estresse>, ou <shampoo> e <xampu>, por exemplo. O consultório de dúvidas do Português, o Ciberdúvidas, responde com frequência a questões sobre a ortografia e utilização de palavras estrangeiras[162]. Já em galego, por outro lado, recomenda-se oficialmente uma substituição do estrangeirismo por uma palavra galega, sempre que isso seja possível:

> Como norma xeral, unha forma procedente doutra língua só debe entrar no galego cando sexa tamén novo o concepto que nomea. De non ser así, débese potenciar a denominación propia, independentemente de que estea viva na fala ou de que se recupere do fondo arcaico ou dialectal. Aplicando estes princípios podemos desbotar formas como <*affaire>,

---

[160] Sobre a influência do castelhano no léxico e morfossintaxe do galego, veja-se Soto & Vidal, 1997.

[161] Para um estudo sobre as diferentes influências de léxico estrangeiro no galego, veja-se Freixeiro Mato (2006: 314-318).

[162] Em http://ciberduvidas.sapo.pt/search.php?keyword=estrangeirismo (23-12-2007) veja-se os vários resultados da procura pelo termo "estrangeirismo" no Ciberdúvidas da Língua Portuguesa. As perguntas chegam de Portugal e do Brasil e andam em torno ora da ortografia, ora da pronúncia do estrangeirismo, ora da pertinência da utilização de um estrangeirismo perante a existência de termo em português. O Ciberdúvidas presta um serviço muito importante à comunidade lusófona e é patrocinado pelo Ministério da Educação português, o que lhe confere já um certo carácter oficial. De qualquer modo, e apesar de defender sempre que possível o uso de um termo português em vez de um estrangeirismo (veja-se a questão acerca do uso de <newsletter> em: http://ciberduvidas.sapo.pt/pergunta.php?id=20999, 23-12-2007), limita-se a responder a dúvidas e não a elaborar reformas ou planificações linguísticas.

> <*chef>, ou <*hall> que suplantan a <asunto>, <xefe de cociña> ou <vestíbulo> respectivamente (…). (Rodríguez Río, 2004)

Também Alcalá (2006) fala do uso e abuso de estrangeirismos de origem anglo-saxónica no âmbito das tecnologias de informação e comunicação e defende a sua substituição, sempre que possível, por uma palavra galega.

Em relação à incorporação de léxico do âmbito científico e técnico no galego, as Normas ortográficas e morofolóxicas do idioma galego recomendam que se siga a forma portuguesa sempre que possível:

> Para o arrequecemento do léxico culto, nomeadamente no referido aos ámbitos científico e técnico, o português será considerado recurso fundamental, sempre que esta adopción non for contraria ás características estruturais do galego. (Normas ortográficas e morfolóxicas do idioma galego, 2003: 12)

Uma adaptação do leitor de estrangeirismos ao PB e ao galego pressupõe a criação de dicionários fonéticos, à semelhança da que foi feito para o PE, a partir dos quais se possam observar tendências. A ausência de recursos desta natureza impediu-nos de prosseguir com a adaptação deste módulo ao PB e ao galego.

De qualquer modo, partindo da análise das transcrições fonéticas de um léxico de cerca de 1400 estrangeirismos, desenvolvido por Cirineu Stein, em trabalho não publicado, apresentamos os principais fenómenos de adaptação fonética imediata dos estrangeirismos no português do Brasil:

- Simplificação das consoantes duplas (ex. <Office> →[Ofisi], <thriller> →[trileX]);
- Palatalização e africamento das consoantes oclusivas alveolares [t] e [d] antes de <i> realizado como [i] em inglês[163] (ex. <ticket> → [tSikEtS][164]; <disk> → [dZisk]);
- Inserção de um shwa epentético entre sequências de consoantes não comuns em Português (ex. <deadline> → [dEdZilajni]);
- Palatalização e africamento de consoante oclusiva alveolar [t] ou [d] em posição final de palavra e inserção de um shwa [i] (ex. <twist> → [twistSi]; <quid> → [kwidZi]);
- Ditongação da sequência do Inglês vogal + <m> em posição final de palavra (ex. <system> → [siste~j~]; <telecom> → [tElEko~w~]);
- Fechamento e arredondamento da vogal final <o> do Inglês (ex. <techno> [tEknu]);
- Manutenção da consoante fricativa <h> do inglês, mas com deslocação do seu ponto de articulação, ou seja, de glotal do Inglês para velar em português do Brasil (ex. <hardware> → [XaXdwEr]; <hobby> → [XObi]);

---

[163] Se em inglês [i] seguido de [t] ou [d] se realizar como [aj], não ocorre este fenómeno: <dial-up> →[dajwap], <timeout> →[tajmawtSi].

[164] Por ausência de estudos, não registamos a marcação de vogal tónica nem a divisão silábica nestas transcrições.

- Em palavra ou locução latina ou francesa, <h> não se lê, tal como na língua de origem (<habeas corpus> → [abeaskOrpus]; <habitué> → [abitue]);
- Semivocalização e nasalização da sequência gráfica <ing> em posição final do Inglês (ex. <marketing> → [maXketSi~j~]);
- Avanço da vogal fechada em posição final de palavra <e> em galicismos (ex. <mademoiselle> → [m6demwazEli]);
- Mudança do ponto de articulação da fricativa interdental inglesa em posição final de palavra para fricativa alveolar surda em PB (ex. <bluetooth> → [blutus]);
- Substituição da fricativa interdental inglesa pela oclusiva dental surda em início de palvra (ex. <thriller> → [tril6X]);
- Paragoge de um shwa [i] em:
    - anglicismos terminados por consoante oclusiva surda [p], [t], [k] (ex. <top>→[tOpi]; <tookit>→[tuwkitSi]; <network> → [nEtwoXki]; <symantec> → [sim6~tEki]);
    - latinismos terminados por consoante oclusiva (ex. <quod> → [kwOdZi]);
    - anglicismos terminados por <e> (ex. <office> → [Ofisi], <deadline> → [dEdZilajni]);
- Mapeamento do vocalismo tónico e átono do Inglês (ex. <Timberlake>→ [tSi~j~berlejki]) e do Francês (ex. <tailleur> → [tajEX]);
- Palatalização da sibilante surda inglesa em início de palavra quando seguida de consoante[165] (ex. <squash> → [iSkwOS]; <slide> → [iSlajdZi]).

Em galego, a renovação lexical através dos estrangeirismos tem sido feita de duas formas: 1) através da *integração do significado*, ou seja, através do "calco"[166], em palavras como <rato> (do inglês <mouse>) ou <rañaceos> (do inglês <sky scrapper>), ou 2) através da *integração do significante*, realizada por duas vias possíveis: quer pelo *estrangeirismo*, "mantendo unha grafía e unha pronuncia estrañas á estructura" do galego[167], quer pela *adaptação* da palavra estrangeira às pautas gráficas e fonéticas da língua receptora (Rodríguez Río, 2004: 412).

---

[165] Este fenómeno não aparece registado neste dicionário fonético de 1400 palavras, que opta por uma articulação mais próxima do original, mantendo a sibilante inglesa, mas surge atestado na ortoépia de Nicola et al. (2003). Trata-se em nosso entender da articulação mais habitual no PB, sobretudo em relação a anglicismos que já entraram completamente na língua.

[166] Freixeiro Mato (2006: 314) define assim o calco: "O calco é un tipo especial de empréstimo en que se imita a significación dunha voz estranxeira, mais non o seu corpo fónico: baloncesto é un calco do ingl. *basquet-ball*, igual que *week-end* xerou o calco fin de semana."

[167] É o caso das designadas "palavras internacionais", em Rodríguez Río (2004: 412) como "hippy", "jazz", "leitmotiv" ou "windsurf", cujas formas originais foram mantidas pela maioria das línguas ocidentais.

Em Rodríguez Río (2003: 107-113) apresentam-se os principais fenómenos de adaptação fonética de estrangeirismos em galego. Uma vez mais não se apresenta informação de transcrição fonética. Porém, é possível descrever os seguintes fenómenos:

- Substituição da sequência gráfica <gn> por <ñ>, mantendo a mesma articulação nasal palatal: <champignon> (do francês <champignon>), <lasaña> (do italiano <lasagna>);
- Simplificação das consoantes duplas: <boicot> (do inglês <boycott>), <gneis> (do alemão <Gneiss>), <bacará> (do francês <baccara>), <tenis> (do inglés <tennis>);
- Grafemas <k> ou sequências <ct> são substituídas em Galego por <c> ou <qu>: <cóctel> (do inglês <cocktail>), <balalaica> (do russo <balalaika>), <quimono> (do japonês <kimono>), <cuscús> (do árabe <kuskus>);
- Ensurdecimento da fricativa palatal estrangeira: <xersei> (do inglês <Jersey>), <xota> (do castelhano <jota>), <beixe> (do francês <beige>);
- Substituição de <sh> por <x>: <xampu> (do inglês <shampoo>);
- Africamento em Galego [tS] da fricativa palatal surda <ch> → [S] do francês: <bechamel>, <chassis>, <fetiche>;
- Realização como fricativa interdental da sequência gráfica <ce>, <ci> que substitui as sequências <ze> e <zi> das línguas de origem: <cebú> (do francês <zebu>, <cinc> (do alemão <Zink>), <gaceta> (do italiano <gazzeta>);
- Subsituição da sequência gráfica <ng> por uma nasal alveolar [n] em anglicismos: <pudin> (de <pudding>), <mitin> (de <meeting>);
- Substituição de <ph> por <f>: <celofán> (do francês <cellophane>), <panfleto> (do inglês <pamphlet>);
- Substituição de <gl> do italiano por <ll>: <pallaso> (<pagliaccio>), <tallarín> (<taglierini>).

No que respeita ao vocalismo, a Figura 45 reproduz as adaptações vocálicas ao galego de galicismos e anglicismos publicada em Rodríguez Río (2003: 109). Pode observar-se que a adaptação se faz no sentido de uma correspondência quase directa.

GALICISMOS

FRANCÉS	GALEGO	EXEMPLOS
-ai- ([ɛ])	*-e-* ([ɛ])	*biais* → ***biés***; *fraise* → ***fresa***; *relais* → ***relé***
-ou- ([u])	*-u-* ([u])	*bouquet* → ***buqué***; *couplet* → ***cuplé***; *sioux* → ***síux***
-au- ([o])	*-o-* ([o], [ɔ])	*landau* → ***landó***; *esquimau* → ***esquimó***; *vol-au-vent* → ***volován***
-eu- ([œ])	*-e-* ([ɛ], [e])	*chauffeur* → ***chofer***
-eau- ([o])	*-o-* ([ɔ])	*fricandeau* → ***fricandó***; *rondeau* → ***rondó***
-ei- ([ɛ])	*-e-* ([ɛ])	*groseille* → ***grosella***
-y- ([i], [j])	*-i-* ([i], [j])	*vichy* → ***vichí***; *boyard* → ***boiardo***
-em-/-en- ([ɑ̃)])	*-am-/-an-*([aŋ])	*ensembler* → ***ensamblar***; *agrément* → ***agremán***
-in- ([ɛ̃])	*-en-* ([ɛŋ])	*meringue* → ***merengue***; *satin* → ***satén***

ANGLICISMOS

INGLÉS	GALEGO	EXEMPLOS
-ee- ([i:]; [i])	*-i-* ([i])	*beefsteak* → ***bisté***; *yankee* → ***ianqui***; *roastbeef* → ***rosbif***
-oo- ([u:]; [u])	*-u-* ([u])	*boomerang*→***búmeran***; *to tatoo*→***tatuar***; *shampoo*→***xampú***
-ai- ([ɛi]; [ɛə])	*-e-* ([ɛ]; [e])	*cocktail* → ***cóctel***; *mohair* → ***moher***
-ea- ([i:])	*-i-* ([i])	*leader* → ***líder***
-y- ([ai])	*-ai-* ([aj])	*nylon* → ***nailon***
-y- ([i]; [j])	*-i-* ([i]; [j])	*dandy* → ***dandi***; *yankee* → ***ianqui***; *penalty* → ***penalti***

Figura 45: Adaptação ao galego do vocalismo de anglicismos e galicismos (Rodríguez Río, 2003:109)

Como trabalho futuro, prevê-se o desenvolvimento de um dicionário fonético de estrangeirismos para o galego, a partir do qual se possa descrever mais fenómenos de adaptação fonética e consequentemente desenvolver um leitor de estrangeirismos mais adequado.

4.5. Síntese do capítulo 4

Os principais tópicos deste capítulo podem resumir-se nos seguintes pontos:

- A leitura de estrangeirismos representa um dos problemas de mais difícil solução para a síntese da fala em português, representando 1,4% dos erros em léxico e 1,3% dos erros em textos;
- Ao nível da síntese da fala em português, o assunto da leitura de estrangeirismos tem merecido pouca atenção;
- A nossa proposta assenta na identificação da língua de origem e passa por três fases: o pré-processamento, que inclui um dicionário com transcrição fonética de palavras que não provêm do inglês ou do francês, o identificador de galicismos e anglicismos (necessário por representarem o maior número de estrangeirismos em português europeu), e dois conversores grafema-fone, um para galicismos e outro para anglicismos;

- As principais dificuldades deste trabalho foram: a dificuldade de obtenção de corpora de análise e a inexistência de reflexão teórica e de normalização no que respeita ao comportamento fonológico e fonético dos estrangeirismos em PE, o que provoca muitas dúvidas de transcrição fonética;
- O leitor de estrangeirismos proposto neste trabalho foi implementado e testado, tendo-se obtido 88,05% de taxa de acerto por palavra e 98,14% de taxa de acerto por fone;
- A ausência de recursos de linguísticos impediu-nos de prosseguir com a adaptação deste módulo ao PB e ao galego; no entanto foram identificados alguns dos fenómenos mais representativos de adaptação fonética imediata dos estrangeirismos em PB em galego;
- Como trabalho futuro é nosso objectivo tratar a marcação de acento tónico e a divisão silábica dos estrangeirismos de forma automática, bem como desenvolver um dicionário fonético de estrangeirismos para o galego.

Capítulo 5

Conversor grafema-fone

A questão da conversão grafema-fone é um assunto que está longe de estar resolvido, como se poderá verificar pelo grande número de publicações sobre o tema, quer em Portugal (Caseiro & Trancoso, 2002; Caseiro *et al.*, 2003; Oliveira *et al.*, 2004; Paiva *et al.*, 2005; Oliveira *et al.*, 2005; Teixeira *et al.*, 2006a; Teixeira *et al.*, 2006b; Oliveira, 2007), quer no Brasil (Seara *et al.*, 2001; Seara *et al.*, 2002; Barbosa *et al.*, 2003a; Seara Jr. *et al.*, 2004; Silva *et al.*, 2006). Apesar de alguns autores considerarem a conversão grafema-fone em sentido estrito, ou seja, apenas a tarefa de transformação de letras em símbolos fonéticos, entendemos este módulo num sentido mais amplo, no qual incluímos o divisor silábico, o marcador de sílaba tónica e o transcritor fonético, visto que as informações dos dois primeiros têm influência na construção do último. Os vários módulos deste capítulo, bem como as suas aplicações ao português do Brasil e ao galego, deram origem a várias publicações, nomeadamente Braga *et al.* (2006a), Braga *et al.* (2006b), Braga *et al.* (2006c), Braga *et al.* (2006d), Braga & Freixeiro (2006), Silva *et al.* (2006) e Braga *et al.* (2007b).

Neste capítulo, propomos algoritmos baseados em regras linguísticas para resolução da problemática da conversão grafema-fone em português. Testes efectuados a cada módulo revelaram as seguintes taxas de acerto: 99,06% para o divisor silábico, 99,54 % para o marcador de sílaba tónica e 99,11% para o transcritor grafema-fone. Os resultados foram discutidos, bem como a sua aplicabilidade ao português do Brasil e ao galego.

5.1. Divisor silábico

5.1.1. Estado da arte

O número de trabalhos em que se reportam divisores silábicos para o PE e para o PB permite constatar que se trata de uma questão da maior importância para o desenvolvimento da naturalidade da fala sintética. De entre os principais trabalhos em que se descrevem divisores silábicos aplicáveis a sistemas de TTS, destacamos as propostas de Teixeira *et al.* (2000) e Teixeira (2004), Seara Jr. *et al.* (2004), Teixeira *et al.*, (2006b) e Oliveira et al. (2005, 2007). Para um bom estado da arte sobre o

trabalho prévio em divisão silábica, veja-se Oliveira *et al.*, 2007. Em todos os casos, os autores reconhecem a importância da identificação da unidade silábica, quer para a implementação de algumas regras do conversor grafema-fonema, quer para a modelização da prosódia, ao nível da duração, intensidade e mesmo frequência fundamental. Em todos os casos, excepto em Barros & Weiss (2006), em que se optou por uma abordagem por máxima entropia, seguiu-se uma abordagem linguística, ora usando algoritmos de processamento da linguagem natural, ora usando máquinas de estados finitos.

Assim, tendo em conta as experiências anteriores mencionadas, desenvolvemos um módulo de divisão silábica de base ortográfica com objectivos fonológicos, ou seja, tentando conciliar as modernas teorias decorrentes da Fonologia (Mateus & Andrade, 2000) com as necessidades práticas subjacentes ao desenvolvimento tecnológico dos sistemas de síntese da fala para o português.

5.1.2. Algoritmos de divisão silábica

Tabela 46: Simbologia usada no algoritmo de divisão silábica.

símbolo	significado
V	vogal (a,e,o, á,é,ó, ú, í, ã, õ, â, ê, ô, à)
G	glide (i, u)
C	qualquer consoante (<lh>, <nh>, CO, CF, CL, CN)
CO	consoante oclusiva (p, t, c+a,o,u; qu+e,i, b, d, g+a,o,u, gu+e,i)
CF	Consoante fricativa (f,v, s, c+e,i, ç, z, ss, ch, j, g+e,i, x)
CL	consoante líquida (l, r, rr excepto <lh>)
CN	consoante nasal (m, n)
SP	espaço
^(+1)= C	o primeiro grafema à direita da vogal é uma consoante qualquer
^(+2) = G	o segundo grafema à direita da vogal é uma glide
^(+3) = V	o terceiro grafema à direita da vogal é uma vogal
^(-1) =CO	o primeiro grafema à direita da vogal é uma consoante oclusiva
→	então

Foram consideradas três condições prévias: 1) uma sílaba nunca pode corresponder a apenas uma consoante; 2) separam-se sempre os grafemas <xc>, <cc>, <ct>, <cç>, <pt> (ex. e<u>xc</u>luir, proje<u>ct</u>o, a<u>cç</u>ão, a<u>pt</u>o); 3) são tratados como um só grafema os dígrafos <nh>, <lh>, <ch>, < rr>, <ss> e os dígrafos <qu> e <gu> quando seguidos de <e> e de <i>.

O algoritmo de divisão silábica assenta numa busca feita por vogal dentro de cada palavra. Na Tabela 46, pode ver-se a simbologia utilizada. Uma vez identificada a vogal, analisa-se a vizinhança grafemática à esquerda e à direita e, de acordo com as regras listadas na Tabela 48, aplicam-se os casos ou saídas apresentados na Tabela 47. Foram consideradas uma sílaba fonológica as unidades constituídas por semivogal seguida de vogal, em <polí<u>ci</u>a> e <ó<u>ci</u>o>, <L<u>ua</u>nda>, <fis<u>io</u>nomia>, <petról<u>eo</u>>, apesar de se considerarem duas sílabas ortográficas.

Tabela 47: Casos e operações considerados.

Caso	Operação
Caso 1	V separa-se do grafema seguinte
Caso 2	V junta-se ao primeiro grafema da direita e separa-se dos grafemas subsequentes
Caso 3	V junta-se ao grafema anterior e separa-se dos seguintes
Caso 4	V junta-se ao grafema anterior e ao seguinte e separa-se dos subsequentes
Caso 5	V junta-se aos dois grafemas seguintes e separa-se do terceiro
Caso 6	V junta-se ao grafema anterior e a todos os grafemas seguintes até final da palavra
Caso 7	V junta-se ao grafema anterior e aos dois seguintes e separa-se dos subsequentes
Caso 8	V junta-se aos dois grafemas anteriores e separa-se do seguinte

Tabela 48: Regras de divisão silábica (versão de 2-08-2007).[168]

#	Regra	Exemplo
1	Se V é início de sílaba e ^(+1)= V → Caso 1	a.eronave, a.inda
2	Se V é início de sílaba e ^(+1) =C e ^(+2) =C e ^(+3) =CO → Caso 5	obs.tar, ads.trito
3	Se V é início de sílaba e ^(+1)= G, <s>, <r>, <l>, <x>, <c> CN e ^(+2)= C → Caso 2	am.bos, en.te, as.pas, al.tura, ar.gúcia, eu.ropa, as.tral, ex.por, ei.ra, ai.po
4	Se V é início de sílaba e ^(+1) =CO, CF e ^(+2) =CO, CN e ^(+3) =V → Caso 2	op.tar, ad.vogar, ag.nóstico, af.ta
5	Se V é início de sílaba e ^(+1)= C e ^(+2)= V, CL → Caso 1	a.rrendar, a-tlas, a.lho, a.mor, a.clamado, a.florar,
6	Se V não é início de sílaba e ^(-1) =C e (^+1)=C e (^+2) =V → Caso 3	ca.lha, ca.la, me.ta, ca.choeira
7	Se V não é início de sílaba e ^(-1) =C e (^+1) =G e ^(+2)= "r" e ^(+3) C → Caso 3	ca-irmos
8	Se V não é início de sílaba e ^(-2) = CO, CF e ^(-1) = CL e (^+1)=C → Caso 8	pro.duto, democra.cia

[168] A primeira versão deste trabalho foi apresentado no *XXI Encontro da Associação Portuguesa de Linguística, em Coimbra*, 2-4 Outubro de 2006 e posteriormente publicado nas actas da APL. A versão que aqui se publica já sofreu actualizações.

Tabela 48: Regras de divisão silábica (versão de 2-08-2007) (continuação).

9	Se V não é início de sílaba e ^(-1) =C e (^+1) =G e (^+2)= <s> e (^+3) =CO→ Caso 7	claus.tro
10	Se V não é início de sílaba e ^(-1) =C e (^+1) = CN e (^+2)= <s> e (^+3)=CO → Caso 7	demons.tra
11	Se V não é início de sílaba e ^(-1) =C, G e (^+1) =G e (^+2)= C → Caso 4	cai.ro, rai.va, quei.xar, cau.sa
12	Se V não é início de sílaba e ^(-1) =C e (^+1) =G e (^+2)= V ou SP → Caso 4	prai.a, mei.a, sei
13	Se V não é início de sílaba e ^(-2) =C e ^(-1) = G e (^+1) =C e (^+2)= V → Caso 3	pia.da, via.gem, sua.da, Sue.ca
14	Se V não é início de sílaba e ^(-1) =C e (^+1) =CL, CN, <s>, <c> e (^+2)= C e (^+3)=V → Caso 4	car.ta, mal.dade, con.tar, spor.ting, hos.pital, pro.jec.to
15	Se V não é início de sílaba e ^(-1) =C e (^+1) =CL, CN, <i> ou (^+2)= <s>, SP → Caso 6	Ama.ral, sóis, se.jam, ima.gens, mais
16	Se V não é início de sílaba é ^(+1) = V igual → Caso 1	ni.i.lismo, re.estruturar
17	Se V não é início de sílaba e ^(-1) =C e ^(+1)=V e ^(+2)=CN → Caso 3	transe.unte, influ.ência
18	Se V não é início de sílaba e ^(-1) = C e ^(+1)= V → Caso 3	lisbo.eta, Lisbo.a, isra.e.lita, pesso.as, cacho.eira
19	Se V não é início de sílaba e ^(-1)=C e ^(+1)= V e ^(+2)=CN e ^(+3) =C → Caso7	influên.cia
20	Se Vogal não é início de sílaba e ^(+1) =CO e ^(+2) =CL e ^(+3) =V → Caso 3	su.blime, de.clarações
21	Se V não é início de sílaba e é = "i" e ^(-2)=<á,é, í, ó, ú> e ^(-1) = C e ^(+1) = <a>, <o> → Caso 6	polí.cia, ó.cio, secretá.ria,
22	Se V não é início de sílaba e é = "i" e ^(-1) = C e ^(+1) = <a>,< o> e ^(+2) = C, <i> → Caso 3	polici.al, polici.ais, Di.as, ri.os, democraci.a, mananci.al
23	Se V não é início de sílaba e é = <ã>,<õ> e ^(-1) = C e ^(+1) = <o>,<e> e/ou ^(+2)=<s> → Caso 6	ambi.ção, cora.ções,
24	Se nenhum dos casos anteriores se verificar e a palavra terminar, a V forma sílaba com os grafemas que restarem até ao espaço em branco ou sinal de pontuação ou hífen.	cas.to, des.cer

5.2. Marcador de sílaba tónica

5.2.1. Estado da arte

A marcação da sílaba tónica tem impacto em dois módulos do sistema de TTS: ao nível do transcritor grafema-fone, por um lado, na medida em que algumas regras de conversão grafema-fone utilizam a informação da tonicidade da vogal (o símbolo de vogal tónica está contemplado na Tabela 51 que descreve a simbologia utilizada no transcritor grafema-fone) e ao nível do módulo de análise e geração prosódica, por outro, na medida em que a tonicidade está associada ao aumento da frequência fundamental, intensidade e duração da vogal ou sílaba abrangidas. A alternância entre sílabas tónicas e átonas é ainda responsável pelo ritmo e por fenómenos de micro-prosódia.

Tabela 49: Simbologia usada no algoritmo de marcação de sílaba tónica.

símbolo	significado
^(0)	último grafema de uma palavra
^(1)	penúltimo grafema de uma palavra
^(2)	antepenúltimo grafema de uma palavra
^(3)	terceiro grafema a contar do final da palavra
^(4)	quarto grafema a contar do final da palavra
T	posição ocupada pela vogal tónica
T=1	a tónica corresponde ao penúltimo grafema
/	excepto, à excepção
→	então
{x}	grafema x
{ }	espaço
1	marcação de vogal tónica

Contudo, são escassos os trabalhos publicados sobre este assunto em concreto. Em Oliveira *et al.* (1991), refere-se a importância deste módulo e diz-se que a versão do então DIXI usa 18 regras previamente descritas em Andrade & Viana (1985), relatório a que não conseguimos ter acesso[169]. Mais recentemente, destacamos o artigo de Teixeira *et al.* (1998), em que se descreve o algoritmo de marcação de sílaba tónica, construído com apenas três regras[170], seguidas de uma tabela de excepções, e o

[169] Esta publicação não está disponibilizada. Foi solicitado a um dos autores o favor de nos disponibilizar toda a bibliografia que achassem importante relacionada com os vários módulos descritos nesta dissertação. Nunca obtivemos resposta.

[170] "A marcação da sílaba tónica obedece às seguintes três regras com prioridade decrescente: 1- o texto escrito que contenha uma das seguintes letras será convertido com a marca de acentuação da palavra (á, é, í, ó, ú, à, è, ì, ò, ù, ã, õ, â, ê, ô). (...) 2- quando não há nenhuma marca de acentuação nas palavras terminadas por uma das seguintes sequências (al, el, il, ol, ul, ar, er, ir, or, ur, az, ez, iz, oz, uz), é colocada uma marca de acentuação na última sílaba.

artigo de Barros & Weiss (2006), em que se treinou um modelo por máxima entropia a partir de um *corpus* de 4219 palavras anotadas com informação de sílaba tónica. No primeiro trabalho, não são conhecidos os níveis de desempenho do sistema nem as tabelas de excepções. Em relação ao desempenho do método estatístico proposto por Barros & Weiss (2006), a taxa de acerto apresentada é de 85,57%.

A constatação da falta de documentação sobre este tema funcionou como motivação para a produção de uma ferramenta de marcação de tonicidade, em colaboração com o Laboratório de Processamento de Sinais, cuja primeira versão pode ser consultada em Silva *et al.* (2006). Estas regras foram inicialmente aplicadas ao PB com um extracto do Cetem-Folha, tendo sido obtidos 98,58% de percentagem de acerto.

Entretanto, testámos estas regras com um *corpus* de 1000 frases do Cetem-Público para o PE, com vista a analisar a adaptabilidade deste algoritmo a sistemas de conversão texto-fala em PE.

5.2.2. Algoritmos de marcação de sílaba tónica

O algoritmo de marcação de tonicidade proposto foi refinado e adaptado ao PE[171], funcionando agora com 31 regras construídas a partir da análise das sequências ortográficas e do conjunto de regras de acentuação gráfica vigentes (Estrela *et al.*, 2004; Bergstrom & Neves, 1997).

Na Tabela 49, apresenta-se a descrição da simbologia utilizada. Na Tabela 50, podem ver-se as regras de marcação de tonicidade.

O algoritmo começa por verificar se existem as excepções no texto, ou seja, as palavras átonas.

Tabela 50: Tabela de marcação da sílaba tónica (versão de 11-12-2007).

#	Regra	Exemplo
1	Lista de palavras átonas	por, um, se
2	Se existir acento agudo, grave[172], circunflexo ou til, a vogal marcada é tónica. Os acentos circunflexo e agudo têm precedência sobre o til.[173]	órgão, órgãos, bênção, bênçãos

3- quando não existe ainda na palavra nenhuma marca de acentuação, é seguida a regra geral de acentuação na penúltima sílaba." (Teixeira *et al*., 1998).

[171] Retirou-se a regra 11 da versão inicial de Silva *et al*. (2006) relativa ao funcionamento de <porque> em PB, já que esta palavra é sempre oxítona em PB.

[172] Excepto: <àquele, àqueles, àquela, àquelas, àqueloutro, àqueloutra, àqueloutros, àqueloutras>, cujo acento grave não deve ser considerado e devem assim cair nas regras seguintes.

[173] Constitui excepção a esta regra nomes ou adjectivos a que se junta o sufixo <-inho>,<-inha>,<- inhos>,<- inhas>,<-zinho>, <-zinha>,<- zinhas>,<- zinhos>, (ex: pãezinhos, sotãozinho) ou <-mente> (ex: cristãmente), em que a vogal tónica se desloca para a penúltima sílaba, embora psico-cognitivamente se possa considerar que existem dois acentos fonológicos.

Tabela 50: Tabela de marcação da sílaba tónica (versão de 11-12-2007) (continuação).

3	Se a palavra só tiver uma vogal → T= vogal	tem, vem, bem, vi,
4	Se ^(0) = {r, l, z, x} → T = 1	propor, carrossel, rapaz, triplex, juiz
5	Se ^(0) = {m} e ^(1) = {i,o,u} → T = 1	pudim, bombom, comum
6	Se ^(0) = {s} e ^(1) = {n} e ^(2) = {i,o,u} → T = 2	pudins, bombons, comuns
7	Se ^(0) = {i} e ^(1) = {u,} e ^(2) = {q,g} → T = 0	caqui, aqui, sagüi
8	Se ^(0) = {s} e ^(1) = {i} e ^(2) = {u,} e ^(3) = {q,g} → T = 1	caquis, sagüis
9	Se ^(0) = {i,u} e ^(1) é vogal → T = 1	caiu, grau, pneu
10	Se ^(0) = {i,u} e ^(1) não é vogal → T = 0	caju, javali
11	If ^(0) = {s} e ^(1) = {i,u} e ^(2) não é vogal → T = 1	cajus, javalis
12	Se ^(0) = {s} e ^(1) = {i,u} e ^(2) é vogal → T = 2	andais, pauis, graus.
13	Se ^(0) = {e} e ^(1) = {u} e ^(2) = {q, g} e ^(3) é vogal/{u} → T = 3	alambique, Henrique, destaque, obrigue
14	Se ^(0) = {s} e ^(1) = {e} e ^(2) = {u} e ^(3) = {q, g } e ^(4) é vogal/{u} → T = 4	alambiques, Henriques, destaques, obrigues
15	Se ^(0) = {e} e ^(1) = {u} e ^(2) = {q, g} e ^(3)={u} → T = 4	açougue, azougue, tougue
16	Se ^(0) = {s} e ^(1) = {e} e ^(2) = {u} e ^(3) = {q, g} e ^(4)={u} → T = 5	açougues, azougues, tougues
17	Se ^(0) = {e} e ^(1) = {u} e ^(2) = {q, g } e ^(3) ={r} → T = 4	embarque, marque, morgue
18	Se ^(0) = {s} e ^(1) = {e} e ^(2) = {u} e ^(3) = {q, g } e ^(4) ={r} → T = 5	embarques, marques, morgues
19	Se ^(0) = {e} e ^(1) = {u} e ^(2) = {q, g} e ^(3) = {n} → T = 4	sangue, exangue, manque, palanque
20	Se ^(0) = {s} e ^(1) = {e} e ^(2) = {u} e ^(3) = {q, g} e ^(4) = {n} → T = 5	sangues, exangues, manques, palanques
21	Se ^(0), ^(1), ^(2) são vogais, se ^(1) = {i,u} e se ^(3) =consoante, { } → T = 2	meia, seio, apoio, aia, drageia, gaia, papagaio
22	Se ^(0) = {s, m} e ^(1), ^(2), ^(3) são vogais, se ^(2) = {i,u} e se ^(4) =consoante, { } → T = 3	meias, seios, apoios, apoiam, drageias, gaias, papagaios
23	Se ^(0) e ^(3) são vogais, e ^(1) é consoante e ^(2) = {i, u} e ^(4) ≠ vogal/ {u} → T = 3	cadeira, queima, louco, estrangeiro
24	Se ^(0) = {s} e ^(1) e ^(4) são vogais, e ^(2) é consoante e ^(3) = {i, u} e ^(5) ≠ vogal/{u} → T = 4	cadeiras, queimas, loucos, estrangeiros
25	Se ^(0) = {a,e,o} e ^(1) é consoante e ^(2) = {n} e ^(3) = {i,u} e ^(4) é vogal → T = 3	ainda, caindo, fluindo, incluindo, oriundo
26	Se ^(0) = {s} e ^(1) = {a,e,o} e ^(2) é consoante e ^(3) = {n} e ^(4) = {i,u} e ^(5) é vogal → T = 4	oriundos

Tabela 50: Tabela de marcação da sílaba tónica (versão de 11-12-2007) (continuação).

27	Se ^(k)[174] = penúltima vogal e ^(k) = {i,u} e ^(k+1) é vogal e ^(k-1) não é vogal e ^(k+2) não é {q,g} → T = k+1	outro, claustro
28	Se ^(0) = {m} e^(1) = {e} e ^(2) = {u} e ^(3) = {q} → T = 1	quem
29	Se ^(0) = {a,o, e} e ^(1) = {i, u} e ^(2) é consoante ou {u} → T = 1	academia, inicie, assobio, conseguia, continua, rua
30	Se ^(0) = {s, m} e ^(1) = {a,o, e} e ^(2) = {i, u} e ^(3) é consoante ou {u} → T = 2	academias, assobios, conseguias, deveriam, continuam, iniciem
31	Se nenhuma das regras anteriores se verificar → T = penúltima vogal da palavra	casa, homem, guerra

Foram considerados átonos, e portanto, desprovidos de tonicidade, os seguintes vocábulos: 1) os artigos definidos <o, a, os, as> e os indefinidos <um, uns>; 2) os pronomes pessoais oblíquos <me, te, se, o, a, os, as, lo, la, los, las, no, na, nos, nas, lhe, lhes, nos, vos> e suas contracções <mo, ma, mos, mas, to, ta, tos, tas, lho, lha, lhos, lhas, no-lo, no-la, no-los, no-las, vo-lo, vo-la, vo-los, vo-las>; 3) o pronome relativo <que>; 4) as preposições <a, com, de, em, por, sem, sob> e as contracções <do, da, dos, das, ao, à, aos, às, no, na, nos, nas, num, nuns>; 5) e as conjunções <e, mas, nem, ou, que, se>. As regras de tonicidade para os substantivos prevêem uma regra de plural, pelo que aparecem muitas vezes repetidas, considerando apenas mais o morfema de plural.

5.3. Transcritor grafema-fone

5.3.1. Estado da arte

A questão da transcrição grafema-fone é uma questão central em Síntese da Fala, constituindo um problema ainda longe de estar solucionado. Além disso, é o módulo por excelência da análise fonética.

Foram propostos vários quadros teóricos para resolver o problema da conversão grafema-fonema nos sistemas de conversão texto-fala, de entre os quais, destacamos os seguintes: árvores de decisão (Lucassen & Mercer, 1984), árvores de decisão treinadas automaticamente (Black *et al.*, 1988), modelos de "Tabela look-up" (Bosh & Daelemans, 1993), abordagens baseadas em dicionários (Coker *et al.*, 1990), abordagens baseadas em regras linguísticas (Kaplan & Kay, 1994), modelos híbridos (Meng *et al.*, 1994), abordagens por redes neuronais (Sejnowski &. Rosenberg, 1987), abordagens por máquinas de estados finitos (Roche & Schabes, 1995; Paulo, 2005), técnicas por Modelos Escondidos de Marcov (Taylor, 2005) e por modelos estatísticos (Chotimongkol & Black, 2000). Em Demper *et al.* (1998), faz-se uma comparação entre várias técnicas e discutem-se os respectivos resultados.

[174] (k) é uma variável, é um dado grafema.

Uma das técnicas mais utilizadas é a abordagem por dicionário, que consiste numa lista de palavras ou léxico, a que se faz corresponder a respectiva transcrição fonética. Esta técnica tem sido mais aplicada a línguas que não apresentam uma correspondência grafema-fonema unívoca, como é o caso do Inglês ou do Francês. Mas esta abordagem falha drasticamente quando surgem palavras que não constam no dicionário, como neologismos, estrangeirismos, etc.

Outra possibilidade são os sistemas mistos, que podem gerar regras linguísticas ou de análise estatística de padrões ortográficos encontrados nas transcrições fonéticas fornecidas pelos dicionários.

No que respeita à transcrição grafema-fone para o PE, há notícia do uso das seguintes abordagens: 1) por regras linguísticas referidas em Oliveira *et al.* (1991), Oliveira, (1996), Teixeira *et al.* (1998), Teixeira (2004); 2) por redes neuronais (Trancoso *et al*. 1994); 3) por CARTs - *Classification and Regression Trees* (Oliveira *et al.*, 2001); 4) por máquinas de estados finitos (Caseiro&Trancoso, 2002; Caseiro *et al.*, 2003; Oliveira *et al.*, 2004); 5) por "machine learning" (Teixeira *et al.*, 2006a, 2006b) e 6) por máxima entropia (Barros & Weiss, 2006).

Em relação ao português do Brasil, salienta-se a abordagem por regras do grupo do LPS (Barbosa *et al.*, 2003a). Em Violaro *et al.* (1999), há referência a um conversor grafema-fone usando um dicionário e com algumas regras desenvolvido pelo grupo do LAFAPE/IEL, para integração no sintetizador por concatenação Aiuruetê. De todas as abordagens apresentadas, os melhores resultados estão reportados em Barbosa *et al.* (2003a), com uma taxa de acerto ao fone de 98,43%.

Descrições mais detalhadas sobre o trabalho prévio feito neste domínio para o Português podem ser encontradas em Teixeira, J.P. (2004: 43-44) e em Teixeira *et al.* (2006b).

O trabalho que passaremos a descrever teve início durante a nossa passagem pelo Laboratório de Sinais e Sistemas. As bases e formulação teórica deste trabalho podem ser encontradas em Teixeira (2004).

Justificamos a nossa abordagem baseada em regras linguísticas apoiando-nos assim em quatro premissas: primeiro, o português é uma língua com bastantes regularidades fonológicas; segundo, uma abordagem por regras é mais económica em termos de memória computacional do que uma abordagem por dicionário; terceiro, nenhuma abordagem que tenha usado métodos estatísticos ou de "machine learning" mostrou ter um desempenho melhor que 98% de fones correctamente transcritos; por último, uma abordagem por regras é sempre capaz de "ler" uma palavra nova.

Para a construção dos nossos algoritmos, foram considerados os estudos mais recentes em Fonética e Fonologia do português (Mateus *et al.*, 1990; Mateus & Andrade, 2000; Rodrigues, 2003).

5.3.2. Algoritmos de transcrição grafema-fone

Na Tabela 51, apresentam-se as convenções de anotação usadas na descrição dos algoritmos propostos para a transcrição grafema-fone.

A nível da anotação fonética, e como já referido, seguimos o alfabeto SAMPA para o português (vide Tabela 1) com uma extensão (a consoante lateral velarizada [l*] que ocorre na articulação da palavra <sal> em PE), por ser o alfabeto mais adequado do ponto de vista do processamento computacional das línguas.

Tabela 51: Símbolos e convenções de anotação usados no conversor grafema-fone.

Símbolo	Significado
...	Qualquer grafema
< x >	Grafema ou conjunto de grafemas <x>
/ y /	Fonema ou conjunto de fonemas y
,	Separa opções
{ x_1, x_2, x_3 }	Conjunto de grafemas
< x_1 {x_2, x_3} >	< x_1 x_2 > ou < x_1 x_3 >
< C / y >	Consoante excepto <y>
< C / {w, z} >	Consoante excepto < w> e < z>
V	Qualquer vogal gráfica (e.g. a, e, i, o, u)
C	Quelquer consoante gráfica (e.g. p, t, k, b, d, g...)
Pont	Sinal de pontuação (e.g. , . ! ? () - ; sp)
Ltr	Caracteres que são letras (e.g. a, b, c, …)
SP	Espaço entre palavras
Hf	Hífen
<(case) x >	Caso que modifica o grafema <x>
<(C) x>	<x> é uma consoante
<(V) x>	<x> é uma vogal
<(UV) x>	<x> é não vozeado
<(VO) x>	<x> é vozeado
<(US) x>	<x> é vogal átona
<(S) x>	<x> é vogal tónica
<(W_bgn) x>	<x> está em início de palavra
<(Prn_M) x>	O grafema <x> é V_S no pron./det. masculino: este(s), esse(s), aquele(s), dele(s)
<(Prn_F) x>	O grafema <x> é V_S no pron./det. feminino: esta(s), essa(s), aquela(s), dela(s)

A nossa transcrição é fonética, e não fonológica, de forma a traduzir com rigor os fenómenos de sandhi, que não são descritos fonologicamente, e procura descrever os hábitos articulatórios do que se considera ser a variedade padrão do português europeu (Rodrigues, 2003; Segura & Saramago, 2001; Veloso, 1999; Ferreira *et al.*, 1995; Cintra, 1995).

O algoritmo começa por fazer um pré-processamento de prefixos gregos e latinos (Tabela 52) e de outras palavras mais frequentes (como <até>, <hoje>, <desde>, <era>) cujo vocalismo não seja previsível por regras. Em seguida, processa as excepções às regras, descritas em rodapé nesta secção.

Tabela 52: Prefixos gregos e latinos no PE.

Prefixo grego/latino	Transcrição fonética	Exemplo
…<aero><-,Ltr>…	[6ErO]	aeronave, aeromotor
…<auto><-,Ltr>…	[awtO]	auto-crítica
…<cardi><-,Ltr>…	[kardi]	cardiologia
…<hemo><-,Ltr>…	[emO]	hemodiálise
…<hepta><-,Ltr>…	[Ept6]	heptassílabo
…<hiper><-,Ltr,sp>…	[ipEr]	hiper-mercado, hipermercado
…<hipo><-,Ltr>…	[‘ipO]	hipocrisia, hipotenusa
…<homo><-,Ltr>…	[OmO]	homofobia, homogéneo
…<macro><-,Ltr>…	[makrO]	macrobiótico
…<mega><-,Ltr,sp>…	[mEg6]	megalómano
…<meso><-,Ltr>…	[mEzO]	mesóclise
…<oftalm><-,Ltr>…	[Oftal*m]	oftalmologia
…<repub><-,Ltr>…	[REpub]	república, republicano
…<retro><-,Ltr>…	[REtrO]	retrospectiva
…<rino><-,Ltr>…	[RinO]	rinoceronte
…<super><-,Ltr,sp>…	[supEr]	super-homem
…<vídeo><-,Ltr>…	[vidjO]	vídeo-conferência

Tabela 53: Regras de transcrição para o grafema <a> do PE.

#	Padrão gráfico de <a>	Fone	Exemplo
1	...<(Rad_G) a>...	[a]	radioterapia
2	... < á, à >...	[a]	rápido
3	... <ão>...	[6~w~]	chão, leilão
4	...<ã>...	[6~]	romã, irmã
5	....<â {m,n}><C/h>...	[6~]	lâmpada
6	... <â> <C>...	[6]	câmara
7	... <am><Pont> ...	[6~w~]	sejam, andam
8	....<a (m,n)><C/h>...	[6~]	campo, canto
9	...<a><l><C/h>...	[a]	calmo, palco
10	...<a><i,u,o><C,Pont>...	[a]	paisagem, ao
11	...<(S) a> <r><Pont>…	[a]	matar, andar
12	...<a> <ct, cç, pt, cc> ...	[a]	acção, captura, faccioso
13	...< (S) a><m,n>...	[6]	ramo, banha
14	...<(S) a>...	[a][175]	cacto, gato
15	... < a >...	[6]	amador

[175] As preposições <para> e <cada> são excepção a esta regra.

Tabela 54: Regras de transcrição para os grafemas <b, c, d> do PE.

#	Padrão gráfico de <b>	Fone	Exemplo
1	... <b>...	[b]	al<u>b</u>atroz
#	**Padrão gráfico de <c>**	**Fone**	**Exemplo**
1	...<c><t,ç>...	[ ][176],[177]	ca<u>c</u>to, a<u>ç</u>ção
2	... < c > < e, i >...	[s]	a<u>c</u>eitar
3	...<cc>...	[ks]	o<u>cc</u>itano
4	... < ç >...	[s]	almo<u>ç</u>o
5	... < ch >...	[S]	a<u>ch</u>o, <u>ch</u>uva
6	... <c>...	[k]	<u>c</u>laro
#	**Padrão gráfico de <d>**	**Fone**	**Exemplo**
1	... <d>...	[d]	<u>d</u>ote, <u>d</u>oa<u>d</u>or

Nas Tabelas 53 a 63 são apresentados os algoritmos de transcrição grafema-fone propostos para o PE, ilustrado com exemplos. A primeira versão destes algoritmos foi publicada em Braga *et al.* (2006a). Seguiram-se novas versões em Braga *et al.* (2006b) e Braga *et al.* (2007b).

Tabela 55: Regras de transcrição para o grafema <e> do PE.

#	Padrão gráfico de <e>	Fone	Exemplo
1	... <SP> <e> <SP>...	[i]	Zé <u>e</u> Ana
2	...< (Prn_M) (S) e >...	[e]	<u>e</u>le, <u>e</u>ste
3	...<(Prn_F)(S) e>...	[E]	<u>e</u>la, <u>e</u>sta
4	...< (Rad_G) e >...	[E]	hip<u>e</u>rtrofia
5	... <õ, ã><e>...	[j~]	põ<u>e</u>s, mã<u>e</u>s
6	...<a><e>...	[j]	Ca<u>e</u>tano
7	...<ém, em><Pont>...	[6~j~]	alguém
8	…<en><s><Pont>…	[6~j~]	imagens
9	... < é >...	[E]	época
10	...< ê, e> <x> <C>....	[6j]	êxtase
11	....<êm ><Pont>...	[6~j~6~j~]	têm, vêm
12	... <ê {m,n}> < C /h >...	[e~]	ciência
13	... < ê >...	[e]	português

[176] Verifica-se excepção a esta regra em algumas palavras em que <c> no mesmo contexto é articulado como [k]: <bráctea, dicção, facção, facto, ficção, fictício, pictórico, secção, sucção>.

[177] [ ] é um fone que não é pronunciado.

Tabela 55: Regras de transcrição para o grafema <e> do PE (continuação) .

14	...<e><i>...	[6]	aceitar
15	...<e><(ct, cç, cc, gn, pç, pt)>...	[E]	dialecto direcção
16	...<e> <n> <Pont>...	[E]	líquen
17	...<e><m,n><e><Pont>..	[E]	gene, leme
18	...<e {m,n}>< C/h>...	[e~]	lento
19	...<e> <sa, se, ssa, za> <s, Pont>...	[e]	chinesa
20	...<e><la><Pont>...	[E][178]	vela, bela
21	...<e><l><C/h, Pont>...	[E]	sensível
22	...<(W_bgn) e> <s> <C>....	[@]	estrada, esperto
23	...<(W_bgn) e><C, rr, ss> <V>...	[i]	exacto, errado
24	...<(W_bgn) e><u>....	[e]	europa, eufemismo
25	...<(W_bgn) he> <C> ...	[i]	herói herança
26	... < (S) e > <m, n > <V>...	[e]	tema, pena
27	...<(S) e><lh, nh, ch, j>...	[6]	velho, lenha
28	...<(S) e><r><Pont>...	[e][179]	ser, manter
29	...<(S) e><u>....	[e]	meu, deu
30	... < (S) e >...	[E]	quero
31	...<C><(US) e> <r> <es, Pont>....	[E]	repórter, Hélder
32	...< (US) e> <o, a>...	[j]	área
33	... < (US) e >...	[@]	índice

Os padrões gráficos e grafemas oriundos de palavras estrangeiros, como <k>, <y>, <w>, <ü>, já foram tratados pelo leitor de estrangeirismos. Os ditongos crescentes e decrescentes foram igualmente considerados, embora alguns tenham sido incorporados nas regras de transcrição dos grafemas <i> e <u>. Ao conceber estas regras, tentou-se, sempre que possível, reduzir a sua dependência em relação aos módulos anteriores da divisão silábica e da marcação de tónica.

Para executar a transcrição dos 22 grafemas do alfabeto português europeu, não considerando os grafemas estrangeiros <k, y, w>, foram necessárias 144 regras de conversão para 38 fones.

[178] <pela>, contracção da preposição <por> com o artigo definido <a>, é excepção a esta regra.
[179] Esta regra funciona para verbos como <colher>[e] mas não para o homógrafo <colher> [E]. Foi necessária a intervenção do desambiguador de homógrafos para este caso.

Tabela 56: Regras de transcrição para os grafemas <f, g, h, i, j, l> do PE.

#	Padrão gráfico de <f>	Fone	Exemplo
1	... < f >...	[f]	faca, afiar
#	**Padrão gráfico de <g>**	**Fone**	**Exemplo**
1	... < g > < e, i >...	[Z]	gelo, giro
2	... < g u > < e, i >...	[g][180]	guindaste
3	... < g >...	[g]	garoto, agora
#	**Padrão gráfico de <h>**	**Fone**	**Exemplo**
1	... < h >...	[ ]	hoje, hospital
#	**Padrão gráfico de <i>**	**Fone**	**Exemplo**
1	...< i{m,n}><C/h, Pont>...	[i~]	limbo, sinto, fim
2	...<í{m,n}><**C**/h>...	[i~]	símbolo
3	... <V/{i, u}> <i>...	[j]	coisa, sai
4	...<i e><Pont>...	[i]	superfície
5	... < i, í >...	[i]	líquido, saí
#	**Padrão gráfico de <j>**	**Fone**	**Exemplo**
1	... < j >...	[Z]	jovem
#	**Padrão gráfico de <l>**	**Fone**	**Exemplo**
1	... < l > <C/h, Pont>...	[l*]	calma, vogal
2	... < l SP> <V>...	[l]	sol a sol
3	... < lh >...	[L]	alho, colher
4	...<l>...	[l]	ali

Tabela 57: Regras de transcrição para os grafemas <m, n> do PE.

#	Padrão gráfico de <m>	Fone	Exemplo
1	... < m >...	[m]	mamã
#	**Padrão gráfico de <n>**	**Fone**	**Exemplo**
1	... <n h>...	[J]	ganho
2	... <n>...	[n]	nata, líquen

[180] Excepções a esta transcrição podem ser encontradas nas seguintes palavras em que <gu> se pronuncia [gw]: <aguentar, antiguidade, arguente, arguição, arguido, consanguinidade, contíguo, contiguidade, ensanguentar, exiguidade, exíguo, lingueta, linguiça, linguista, pinguim, sagui, saguim, sanguinário, sanguinolento, unguento, unguiforme>.

Tabela 58: Regras de transcrição para o grafema <o> do PE.

#	Padrão gráfico de <o>	Fone	Exemplo
1	<Pont><o><Pont>	[u]	o bolo é bom
2	...< (Rad_G) o >...	[O]	hemoglobina
3	....<ó>...	[O]	código, avó
4	... < õ >...	[o~]	corações, põe
5	...<o{m,n}><C/h, Pont>...	[o~]	compor, conde, som
6	... <ô n ><C/h>...	[o~]	gôndola
7	... < ô >...	[o]	sôfrego, pôr
8	...<ou>...	[o]	ouvir, couve
9	...<o><i>...	[o]	dois, oito
10	...<(S) o><r><es, Pont>...	[o][181]	compor, dor, sabor, sabores
11	...<o><z><Pont>...	[O][182]	voz, atroz
12	...<o><so><Pont>...	[o]	saudoso, caprichoso
13	...<o><sa, sos, sas> <Pont>...	[O]	saudosa, virtuosa
14	...<(W_bgn) o><l>...	[O]	olá, olhar
15	...<Ltr><o><l><C/h>...	[o]	soltar, voltar
16	...<o><l><Pont>...	[O]	futebol
17	...< (W_bgn) h> <o> <r, s, t>...	[O]	hortelã, hora, hostil, hotel
18	... <(S) o > <m, n>...	[o]	soma, sono
19	...<(S) o><a>...	[o]	perdoa, canoa
20	...<(S) o><o><Pont>...	[o]	voo, enjoo
21	...<a><o><C/n, Pont>...	[w]	ao, caos
22	...<(W_bgn)o><C><C>...	[O]	oclusão, obtuso
23	... <(US) o >...	[u]	carros, opor
24	...< (S) o >...	[O]	embora, agora

As regras de transcrição do grafema <o> são completadas com uma lista de substantivos cuja vogal tónica é semi-fechada no singular (ex. <corvo> [o]) mas semi-aberta no plural (ex. <corvos> [O]) e até no feminino, no caso de <sogra> e <nova>. Uma vez que, de acordo com as regras apresentadas na tabela de conversão do grafema <o>, o output do algoritmo para estes casos é [O], a Tabela 59 contém apenas o singular destas palavras. Na Tabela 59, apresentam-se ainda os substantivos que conservam a vogal tónica semi-fechada no singular e no plural (Cunha & Cintra,

[181] Excepções a esta regra encontram-se nas seguintes palavras e suas variações morfológicas, em que <o> se pronuncia [O]: <maior, menor, melhor, pior, suor, sénior, júnior>.
[182] A palavra <arroz> é excepção: <o> articula-se [o].

1992). Muitas destas palavras são homógrafos, pelo que são previamente processadas pelo desambiguador de homógrafos.

Tabela 59: Substantivos cuja vogal tónica é [o].

Substantivos cuja vogal tónica é [o] apenas no singular	abrolho, caroço, contorno, corcovo, coro, corno, corpo, corvo, despojo, destroço, escolho, esforço, estorvo, fogo, forno, foro, fosso, imposto, jogo, miolo, novo, olho, osso, ovo, poço, porco, porto, posto, povo, reforço, rogo, sobrolho, socorro, sogro, tijolo, toco, tojo, tordo, torno, troço, troco
Substantivos cuja vogal tónica é [o] no singular e no plural	acordo, adorno, bojo, bolo, cachorro, coco, colmo, consolo, dorso, encosto, engodo, estojo, ferrolho, globo, golfo, gosto, lobo, logro, moço, morro, mosto, namoro, piloto, piolho, poldro, polvo, potro, reboco, repolho, restolho, rolo, rosto, sopro, suborno, topo

Tabela 60: Regras de transcrição para os grafemas <p, q, r> do PE.

#	Padrão gráfico de <p>	Fone	Exemplo
1	...<p><t, ç>...	[ ][183]	óptimo
2	... < p >...	[p]	pato
#	**Padrão gráfico de <q>**	**Fone**	**Exemplo**
1	... < q u >< i, e> ...	[k][184]	quilo, quente
2	... < q >...	[k]	qual, quorum
#	**Padrão gráfico de <r>**	**Fone**	**Exemplo**
1	... < r r > ...	[R]	carro, arrendar
2	...<(W_bgn) r>...	[R]	rua, rio, rocha
3	... < r >...	[r]	mar, cara

[183] em curso a construção de uma lista em que <p> se articula: <helicóptero, aptidão>.

[184] Excepções a esta transcrição podem ser encontradas nas seguintes palavras em que <qu> se pronuncia [qw]: <aquícola, aquista, cinquenta, consequência, delinquência, delinquir, deliquescência, eloquência, eloquente, equestre, equidade, equídeo, equidistante, equitativo, exequível, frequência, frequente, obliquidade, querco, quercite, quingentésimo, quinquagenário, quinquagésimo, quiproquó, sequela, tranquila, tranquilidade, ubiquidade>.

Tabela 61: Regras de transcrição para os grafemas <s, t> do PE.

#	Padrão gráfico de <s>	Fone	Exemplo
1	...<(W_bgn) s>...	[s]	saúde, sim
2	... <V> <s> <V>...	[z]	asa
3	...<s><SP><C_VO>...	[Z]	olhos verdes
4	...<s><SP><C_UV>...	[S]	olhos castanhos
5	...<s><SP><V, h>...	[z]	os olhos
6	... < s s > ...	[s]	assar
7	... <s><C_VO>...	[Z]	rasgar
8	... <s><C_UV, Pont/SP>...	[S]	rasca, olhos
9	... <tr(a, â) n> <s> <V>...	[z]	transitar
10	... <ob> <s> <éq>...	[z]	obséquio
11	...<s>...	[s]	cansado
#	**Padrão gráfico de <t>**	**Fone**	**Exemplo**
1	... <t>...	[t]	tacto

Tabela 62: Regras de transcrição para os grafemas <u, v> do PE.

#	Padrão gráfico de <u>	Fone	Exemplo
1	...<m><ui><t>...	[u~j~]	muito
2	... < ü > ...	[w]	linguística
3	... <g, q> <u><a, o>...	[w]	qual, quorum
4	... <u {m, n} > <C /h, Pont>...	[u~]	abundante, retumbante
5	... <é, e, a, i>< u >...	[w]	céu, seu, caudal
6	... < ú, u >...	[u]	acústica
#	**Padrão gráfico de <v>**	**Fone**	**Exemplo**
1	... < v >...	[v]	voando

A nível da implementação dos algoritmos tentou-se, sempre que possível, que as regras que convergissem para a saída *default* não fossem implementadas. O *default* é considerado a saída mais frequente no total de todas as regras propostas para um dado grafema. Os latinismos são processados nesta fase e não no módulo de estrangeirismos, dado que sofrem uma adaptação fonética profunda, o que permite mapear o sistema fonológico original com o sistema fonológico do português.

Tabela 63: Regras de transcrição para os grafemas <x, z> do PE.

#	Padrão gráfico de <x>	Fone	Exemplo
1	... <(e,ê)> <x> <C_UV>...	[S]	ê<u>x</u>tase, te<u>x</u>to
2	...<(W_bgn) ine> <x><V>...	[z]	ine<u>x</u>orável, ine<u>x</u>istente
3	... <(W_bgn) e><x><V>...	[z]	e<u>x</u>ecrar, e<u>x</u>ame, e<u>x</u>agero
4	... <(W_bgn) e> <x> <Hf> <C_VO>...	[Z][185]	e<u>x</u>-marido
5	...<(W_bgn)e> <x> <Hf><C_UV>...	[S]	e<u>x</u>-padre
6	... <(W_bgn) e> <x> <Hf> <V>...	[z]	e<u>x</u>-aluno
7	... <(W_bgn) x>...	[S]	<u>x</u>adrês, <u>x</u>aile
8	...<trou><x>...	[s]	trou<u>x</u>e
9	...<m, pr><o, ó, a, á><x><im>...	[s]	má<u>x</u>imo, pro<u>x</u>imidade
10	...<au><x><íl, il>...	[s]	au<u>x</u>ílio
11	....<fl><e,u><x>...	[ks][186]	fle<u>x</u>ão, reflexo
12	...<ne, fi, se><x>...	[ks][187]	ane<u>x</u>o, fi<u>x</u>ar
13	...<x> <Pont>...	[ks][188]	ino<u>x</u>, duple<u>x</u>
14	... <x>...	[S][189]	en<u>x</u>ofre
#	**Padrão gráfico de <z>**	**Fone**	**Exemplo**
1	... < z SP> <C_UV>...	[S]	fa<u>z</u> favor
2	... < z SP> <C_VO>...	[Z]	fa<u>z</u> bem
3	... < z SP> < V, h>...	[z]	fa<u>z</u> anos
4	... < z > < Pont/ SP>...	[S]	arro<u>z</u>, fa<u>z</u>
5	... < z >...	[z]	<u>z</u>umbido

[185] Excepção a esta regra pode ser encontrada na expressão latina: <ex-libris> [ks].
[186] Excepção a esta regra pode encontrar-se na palavra <reflexão> articulada com [s].
[187] Excepção a esta regra pode encontrar-se na palavra <fixe> articulada com [S].
[188] Excepção a esta regra pode encontrar-se na palavra <cóccix> articulada com [S].
[189] Excepções a esta transcrição podem encontrar-se nas seguintes palavras em que <x> se pronuncia [ks] no meio da palavra: <abnóxio, apoplexia, axial, axila, axiologia, axioma, bissexual, circunflexo, complexão, complexo, conexo, convexão, convexo, crucifixo, filoxera, fixação, fixar, fixo, fluxo, galáxia, heterodoxo, indexação, infixo, inoxidável, intoxicar, íxia, léxico, lexicografia, marxismo, maxilar, maximizar, nexo, nóxio, obnóxio, ortodoxo, oxalato, oxidação, oxidar, oxigénio, oxítono, oxiúro, paradoxo, paralaxe, paroxismo, paroxítono, perplexo, praxis, prefixo, prolixo, proparoxítono, saxofone, sexagésimo, sexagenário>.

5.4. Testes e discussão de resultados

Os testes do divisor silábico, marcador de tonicidade e transcritor grafema-fone (cf. Figuras 46, 47), à semelhança do que se fez para os módulos anteriores, correram na aplicação TTS Front-End. Obtiveram-se assim as seguintes taxas de acerto: 99,44% para o divisor silábico, 99,59% para o marcador de tonicidade e 99,28% para o transcritor grafema-fone.

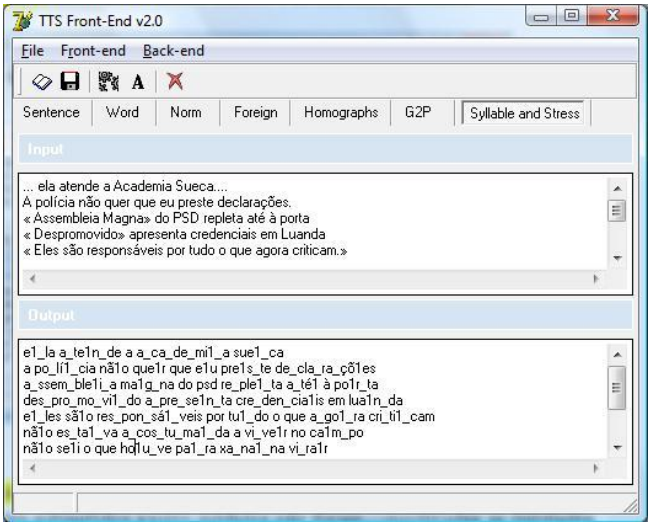

Figura 46: Interface do divisor silábico e do marcador de tónica.

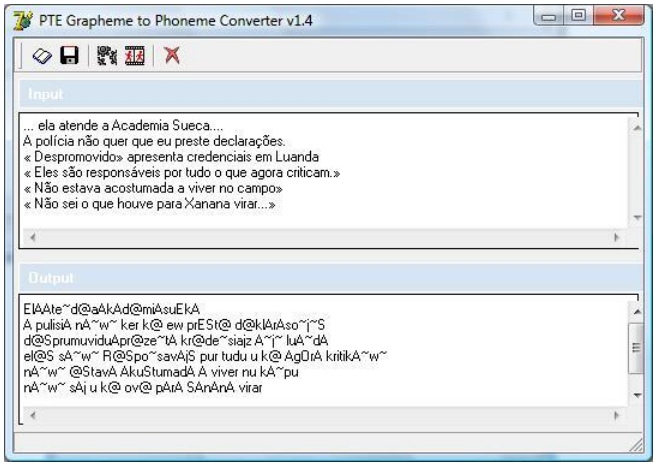

Figura 47: Interface do transcritor grafema-fone.

Nas Tabelas 64, 65 e 66 apresentam-se os erros encontrados em cada um dos módulos. Nos testes submetidos a estes módulos não foram consideradas as entidades

que são processadas no módulo de normalização do texto, como siglas, acrónimos, numerais, abreviaturas, etc, por pertencerem ao nível anterior do processamento de texto. Durante o processo de refinamento de cada um dos módulos foram usadas 500 frases, com 4356 palavras, aleatoriamente escolhidas de um *corpus* constituído por vários tipos de textos, desde jornalístico, a literário, técnico e oral transcrito.

O divisor silábico e o marcador de tonicidade foram testados usando 1000 frases extraídas automaticamente do *corpus* do Cetem-Público[190] contendo 8052 palavras e 41156 caracteres sem espaços.

Tabela 64: Erros resultantes do teste do divisor silábico com frases em PE.

Tipo de erro	# erros	% erros
Estrangeirismos	12	0,15
Não separação de hiatos	33	0,41
Total	**45**	**0,56**

Na Tabela 64, estão registados os erros resultantes do output do divisor silábico testado com as 1000 frases do Cetem-Público. Foram registados 0,56% de erros numa proporção de um erro por palavra. O grande número de erros (0,41%) deve-se à não identificação de certos hiatos, que são entendidos como ditongos e, por consequência, não são separados. Estes casos ocorrem em palavras como <lisb<u>oe</u>ta> (lis_boe1_ta), <r<u>ee</u>screve> (rees_cre_ve), <pr<u>oi</u>bir> (proi_bi1r), <C<u>oi</u>mbra> (co1im_bra), <p<u>oe</u>mas> (poe1_mas). O segundo tipo de erros ocorre em estrangeirismos de diversas origens, sempre que se verifiquem sequências de consoantes estranhas à estrutura silábica do português, como <Frankfurt> (fra1n_kfurt), <Sydney> (syd_ne1y), <requiem> (re_qui1_em), <cocktail> (coc_kta_i1l).

A comparação dos nossos resultados com os valores documentados para sistemas com funções análogas (99,94% em Teixeira *et al.*, 2000 e 99,77% e 99,59% em Oliveira *et al.*, 2005[191] e 97,64% em Barros & Weiss, 2006) deixa-nos optimistas, apesar de sabermos que ainda é necessária alguma refinação no algoritmo.

Ao nível do marcador de tonicidade, o maior número de erros dá-se com palavras estrangeiras. Na Tabela 65, pode ver-se que de 0,41% de palavras cujas vogais tónicas são mal assinaladas, 0,35% dos erros ocorre em estrangeirismos. O grande número de anglicismos na língua justifica o número de erros na marcação de tónica dentro da categoria dos estrangeirismos (vide Figura 48), com 0,11%, em palavras como

[190] Disponível em http://www.linguateca.pt/CETEMPublico/ (12-12-2007).

[191] Os autores apresentam uma comparação de dois métodos de divisão silábica, um por FSTs e outro com um algoritmo baseado na proposta teórica para a silibificação do português de Mateus & Andrade (2000). Ambas as abordagens foram testadas com dois corpora. Os melhores resultados ocorrem com o algoritmo de Mateus & Andrade. 99,77% e 99,59% referem-se aos melhores resultados desse algoritmo em relação às fronteiras de sílabas correctamente marcadas.

<internet> (in_te1r_net) ou <cocktail> (coc_kta_i1l). Com o mesmo número de erros seguem-se as palavras com outras origens (0,11%), onde se encontram <jihad> (ji1_had) ou <Arafat> (a_ra1_fat). Com 0,04% surgem os italianismos, presentes em marcas (<Lamborghini>) e nomes próprios (<Pausini>), ao lado dos galicismos e dos latinismos. Na categoria "outros" incluem-se a palavra <Coimbra> (co1im_bra), repetida 3 vezes, e a organização <Quercus> (quer_cu1s).

Tabela 65: Erros resultantes do teste do marcador de tonicidade com frases em PE.

Tipo de erro	# erros	% erros
Estrangeirismos	28	0,35
Outros	5	0,06
Total	**33**	**0,41**

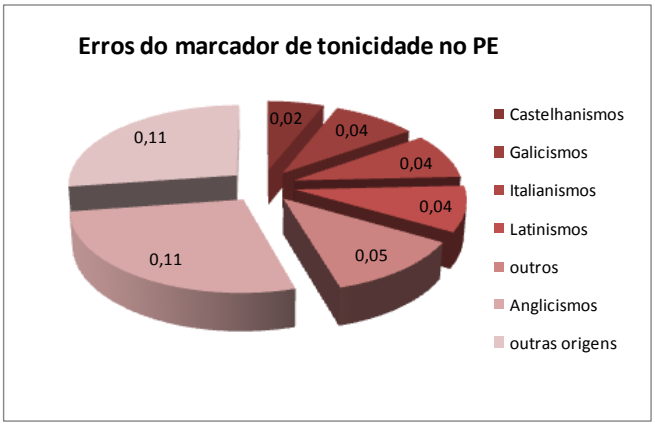

Figura 48: Resultados detalhados dos erros do marcador de tonicidade para PE.

O confronto dos nossos resultados com outras avaliações de marcadores de tonicidade (85,57% em Barros & Weiss) vem demonstrar um melhor desempenho de uma abordagem baseada em regras linguísticas.

No que respeita aos erros encontrados no transcritor grafema-fone, constata-se que a maior parte deles tem a ver com erros na decisão do timbre vocálico dos grafemas <e> (0,27%) e <o> (0,21%). Trata-se essencialmente de erros de transcrição da vogal tónica em nomes (<cabeça>[e], <revolta>[O], <comboio>[O]) e adjectivos (<negro>[e], <fresco>[e]) e em formas verbais, em que é frequente haver alternância da vogal tónica ao longo da flexão verbal (ex. <fez>, <teve>, <bebo> [e], <bebe>[E]). Seguem-se os erros na transcrição da vogal <a>, com 0,09%, que ocorrem muitas vezes em palavras cuja vogal tónica é semi-fechada (ex. <cada>) ou com prefixos ainda não considerados (ex.<extraordinária>, <magnitude>).

Uma vez mais, os estrangeirismos representam uma dificuldade de leitura neste algoritmo, com cerca de 0,08% de erro. Trata-se de nomes próprios estrangeiros (<Carter>, <Arafat>, <Lenine>, <Drosnin>) e de empresas (<Parmalat>).

Outro problema decorre da dificuldade de previsão do timbre dos grafemas <a> e <o> em advérbios de modo derivados de adjectivos (ex: <obrigatoriamente> [O], <delicadamente> [a]), que graficamente a ser átonos, mantendo-se no entanto com o timbre aberto da vogal original. Os advérbios de modo foram considerados à parte, embora na verdade se trate de erros na transcrição de vogais <a> e <o>. A sua contagem em separado deve-se ao facto de ser tratar de um assunto que que necessita de um automatismo de identificação de acento secundário, tal como as palavras derivadas e que, por isso, merecerá atenção em trabalho futuro.

Descrições de outros sistemas similares apresentam resultados menos interessantes (2,08% de erro em Oliveira, 1996; 97,7% de acerto em Oliveira *et al.*, 2004; 88,94% de acerto em Barros & Weiss, 2006; 2,66%-2,91% de erro, os melhores resultados do método "machine-learning", usando a informação silábica, em Teixeira *et al.*, 2006) ou não mencionam avaliação (Teixeira *et al.*, 1998).

Tabela 66: Erros resultantes do teste do transcritor grafema-fone com frases em PE.

Tipo de erro	# erros	% erros
Erros na transcrição de <a>	12	0,09
Erros na transcrição de <e>	36	0,27
Erros na transcrição de <i>	2	0,01
Erros na transcrição de <o>	28	0,21
Estrangeirismos	11	0,08
Nomes próprios	4	0,03
Topónimos	4	0,03
Advérbios em -mente	4	0,03
Total	**174**	**0,82**

Em síntese, os estrangeirismos são responsáveis por grande parte dos erros ao nível da conversão grafema-fone, o que justificou o desenvolvimento de um módulo especializado, o leitor de estrangeirismos, apresentado no Capítulo 4.

5.5. Aplicações do sistema ao português do Brasil

5.5.1 Divisão silábica e marcação da sílaba tónica – testes e resultados

Em Silva *et al.*, 2006, trabalho desenvolvido em parceria com o Laboratório de Processamento de Sinais da UFRJ, apresenta-se o resultado do teste de uma versão

anterior do marcador de tonicidade usando textos em PB, com uma taxa de acerto à palavra de 98,58%.

Posteriormente, após melhoramentos introduzidos nos módulos de divisão silábica e marcação de tonicidade, testaram-se os módulos desenvolvidos para o PE e apresentados nas secções 5.1 e 5.2. em 500 frases extraídas aleatoriamente do Cetem-Folha[192], contendo 5372 palavras e 28633 caracteres sem espaços. Não houve qualquer tipo de adaptação do nosso sistema à variedade do português do Brasil. Os resultados do teste (cf. Tabelas 67 e 68) revelaram que as taxas de acerto do divisor silábico e do marcador de tónica são 99,20% e 99,60% respectivamente.

Tabela 67: Erros resultantes do teste do divisor silábico com frases em PB.

Tipo de erro	# erros	% erros
Estrangeirismos	10	0,18
Problema de identificação de semivogais	32	0,59
Siglas	1	0,01
Total	**43**	**0,80**

Tabela 68: Erros resultantes do teste do marcador de tonicidade com frases em PB.

Tipo de erro	# erros	% erros
Estrangeirismos	17	0,31
Diferenças de posição de acento	4	0,07
Acrónimos	1	0,01
Total	**22**	**0,40**

Em relação ao divisor silábico, os resultados revelam que a maioria dos erros de separação silábica é motivada por dificuldades de separação de ditongos (ex. <funcionários> fun_ci_o_ná1_rios), seguida de palavras estrangeiras (ex. <Wa-shin-gton> was_hi1n_gton; <har-dwa-re> har_dwa1_re). Outro tipo de erros dá-se perante siglas <SOS> (so1s).

Em relação ao marcador de tónica, a maior parte dos erros encontra-se também nos estrangeirismos, que representam 0,31% dos erros (ex. <Corinthians> co_rin_thi1_ans). Segue-se a palavra <porque>, que é oxítona em PB, enquanto que em PE é paroxítona, com 0,07% dos erros e acrónimos, como <Telesp> (te1_lesp).

Estes resultados vêm demonstrar a grande taxa de aplicabilidade dos algoritmos de divisão silábica e marcação de tonicidade do PE ao PB.

[192] Disponível em: http://www.linguateca.pt/CETENFolha/ (12-12-2007).

5.5.2. Transcritor grafema-fone – adaptação, testes e resultados

Neste trabalho, realizado em parceria com o Laboratório de Processamento de Sinais da UFRJ (Silva *et al.*, 2006), pode ver-se o resultado da adaptação ao PB dos algoritmos propostos inicialmente para o PE na secção 5.3. da presente dissertação.

Na Tabela 69, apresenta-se o inventário fonético do PB representado em SAMPA[193] que foi usado na construção das regras de transcrição grafema-fone. Embora não sendo muitas as diferenças, as principais ocorrendo ao nível do vocalismo átono e em relação a certas consoantes[194], o PB requer um alfabeto fonético distinto do PE. Apesar de alguma discussão e polémica em relação ao que se considera a variedade standard do PB, bem mais difícil de definir que a do PE, e apesar do grande prestígio da variedade carioca[195], seguimos para este trabalho aquilo que se considera ser a variedade "neutra", ou seja, a variedade mais próxima da locução dos jornalistas que apresentam o *Jornal Nacional* brasileiro, transmitido pela rede de televisão Globo:

> Many pronunciation features that are unique to carioca dialect are suppressed in the standard broadcast speech that one hears in programs such as O Jornal Nacional, the national news program whose popularity equals that of the novela. (…) Professional broadcasters avoid these patterns for a good reason. Outside of Rio, Brazilians may perceive these traits as affected or pretentious. (Giangola, 2001: 13)

Nas Tabelas 70 a 79 descrevem-se as regras de conversão grafema-fone para o PB. Estas regras sofreram algumas adaptações, mas a maioria dessas alterações está relacionada com o output fonético, que na variedade do PB é por vezes diferente do PE. A simbologia e convenções utilizadas na descrição das regras é a mesma apresentada para o PE na secção 5.3. Para cada grafema, começa-se com as regras que cobrem os casos mais raros e termina-se com uma regra para o *default*. Não foram construídas regras para transcrever o grafema <x>, dada a imprevisibilidade das suas realizações, pelo que a transcrição é feita através de uma tabela de excepções seguida da sua transcrição fonética.

---

[193] Não existe página oficial do SAMPA para português do Brasil, mas esse inventário, construído a partir do SAMPA para PE (http://www.phon.ucl.ac.uk/home/sampa/portug.htm, 13-12-2007), tem sido usado por vários investigadores que trabalham em Síntese da Fala. O inventário completo pode ser encontrado em Barbosa et al. (2003) e em Maia (2006).

[194] O PB apresenta africadas palatais, que desapareceram no PE por volta do século XVIII (Teyssier, 2001: 53-54; Castro, 1991: 258), e diferentes realizações do grafema <r>.

[195] Callou & Leite (2002) invocam razões de ordem histórica para justificar o grande prestígio de que o dialecto falado no Rio de Janeiro ainda goza: o Rio de Janeiro foi capital do Império português entre 1808 e 1821; depois da Independência de 1822, o Rio foi capital do Estado Brasileiro até 1960.

Tabela 69: Alfabeto SAMPA para PB.

classe de fonema	símbolo	exemplo
Vogais orais	**[a]**	jatobá, capacete
	/E/	é, pele, velho
	[e]	capacete, resolver, respeito
	[i]	lápis, justiça, ele
	[O]	ópio, jogos, sozinho, forte
	[o]	jogo, golfinho, corpo
	[u]	Raul, baú, logo
vogais nasais	**[a~]**	andar, canção, cama
	[e~]	então, tempo, menos
	[i~]	ninho, tinta, latina, importa
	[o~]	onda, campeões, somos, homem
	[u~]	um, muito, umbigo
semi-vogais	**[j]**	pai, foi, micróbio
	[w]	fácil, eu, quase
	[j~]	muito, parabéns, compõe
	[w~]	estavam, não
consoantes oclusivas	**[p]**	papai, psicólogo, apto
	[t]	pato, constituinte
	[k]	casca, quero, quanto
	[b]	barba, absinto
	[d]	dados, administrar
	[g]	guerra, gato, agüentar,
consoantes africadas	**[tS]**	tia, pacote, constituinte
	[dZ]	dia, cidade, disco
consoantes fricativas	**[f]**	fanfarrão, afta, afluente
	[s]	sapo, caçar, sessão, lápis
	[S]	chá, xaveco, cachorro
	[v]	vovó, vamos, avião
	[z]	casa, coisa, quase, exato
	[Z]	geladeira, trovejar
	[X]	casar, certo, arpa, arco
consoantes nasais	**[m]**	mamãe, emancipar
	[n]	nome, atenuar,
	[J]	casinha, galinha
consoantes líquidas	**[l]**	laranja, leitão
	[L]	calhar, colheita, melhor
	[R]	carro, rua, rato, carga, germe
	[r]	garoto, frango, por exemplo

Tabela 70: Regras de transcrição para os grafemas <a, b, c, d> do PB.

#	Padrão gráfico de <a>	Fone	Exemplo
1	...<an><Pont>...	[a˜]	Ivan, Itapoan
2	...<am><Pont>...	[a˜w˜]	andam, cresçam
3	...<(S) a><m,n>...	[a˜]	cama, banho
4	...<â(m,n)><C-h>...	[a˜]	lâmpada, cântico
5	...<ão>...	[a˜w˜]	avião
6	...<ã, â>...	[a˜]	amanhã, câmara
7	...<à, á>...	[a]	Antártica, àquela
8	...<a(m,n)><C-h>...	[a˜]	antropofagia, ambos
9	...<a>...	[a]	aracnofobia
#	**Padrão gráfico de <b>**	**Fone**	**Exemplo**
1	...<b>...	[b]	abacate
#	**Padrão gráfico de <c>**	**Fone**	**Exemplo**
1	...<c><e,i>...	[s]	aceitar, jacinto
2	...<ç>...	[s]	almoço
3	...<c h>...	[S]	acho
4	...<c>...	[k]	claro
#	**Padrão gráfico de <d>**	**Fone**	**Exemplo**
1	...<d><i,[i] >...	[dZ]	dia, tarde
2	...<d><C-r,l>...	[dZ]	advogado
3	...<d><Pont>...	[dZ]	raid
4	...<d>...	[d]	dote

Tabela 71: Regras de transcrição para os grafemas <e> do PB.

#	Padrão gráfico de <e>	Fone	Exemplo
1	...<en><Pont>...	[e˜]	hífen, líquen,
2	...<(S)e ><l><C-h,Pont>...	[E]	papel, relva, mel, selva
3	...<(US) e><l><C-h,Pont>...	[e]	sensível, selvagem
4	...<õ,ã><e>...	[j˜]	mães, corações, alemães
5	...<a><e>...	[j]	Caetano
6	...<(Prn M) (S)e>...	[e]	ele, este
7	...<(Prn F) (S)e>...	[E]	ela, esta
8	...<(W bgn) e><x><V>...	[e]	exato, existe
9	...<(W bgn) e,ê><x,s><C>...	[e]	exceto, excelente, êxtase, estrada
10	...<(W bgn) e><C-(m,n,x)>...	[e]	errado, económico
11	...<e><ne,me><Pont>...	[e]	creme, higiene, gene, leme

Tabela 71: Regras de transcrição para os grafemas <e> do PB (continuação).

12	...<e><la,lo><Pont>...	[E][196]	v<u>e</u>la, aquar<u>e</u>la
13	...<e><sa,ssa,za><Pont>...	[e]	chin<u>e</u>sa, cond<u>e</u>ssa, pobr<u>e</u>za,
14	...<ê (m,n)><C-h>...	[e˜]	ci<u>ên</u>cia, exist<u>ên</u>cia
15	...<e (m,n)><C-h>...	[e˜]	<u>em</u>bora, <u>en</u>toação
16	...<(S)e><m,n>...	[e˜]	t<u>e</u>ma, com<u>e</u>mos, c<u>e</u>na, p<u>e</u>na
17	...<(e, é, ê)m><Pont>...	[e˜ j˜]	cont<u>ém</u>, t<u>êm</u>, ont<u>em</u>
18	...<ê>...	[e]	bebê
19	...<é>...	[E]	época
20	...<e><s><Pont>...	[i]	frases
21	...<(S)e i>...	[e j]	fiquei, areia, feio
22	...<e><Ltr><S_V>...	[e]	semestre, perigo
23	...<Pont><e><Pont>...	[i]	o João <u>e</u> a Ana
24	...<e><Pont>...	[i]	índic<u>e</u>
25	...<e>...	[e]	b<u>e</u>bê

Tabela 72: Regras de transcrição para os grafemas <f, g, h, i> do PB.

#	Padrão gráfico de <f>	Fone	Exemplo
1	...<f>...	[f]	<u>f</u>aca
#	**Padrão gráfico de <g>**	**Fone**	**Exemplo**
1	...<g><e,i>...	[Z]	<u>g</u>elo
	...<g u><e,i>...	[g]	<u>gu</u>indaste
	...<g>...	[g]	<u>g</u>aroto
#	**Padrão gráfico de <h>**	**Fone**	**Exemplo**
1	...<h>...	[ ]	<u>h</u>oje
#	**Padrão gráfico de <i>**	**Fone**	**Exemplo**
1	...< (S) u><i><t>...	[j˜]	mu<u>i</u>to
2	...<i e><Pont>...	[i]	superfíc<u>ie</u>, planíc<u>ie</u>
3	...<i,í (m,n)><C-h,Pont>...	[i˜]	s<u>in</u>to, s<u>ím</u>bolo, f<u>im</u>
4	...<i><m,n><V,h>...	[i˜]	<u>i</u>noperante, <u>i</u>maginar, n<u>i</u>nho
5	...<V-i><i>...	[j]	co<u>i</u>sa, sa<u>i</u>,
6	...<i, í>...	[i]	l<u>í</u>quido, sa<u>í</u>

[196] Excepções a esta regra ocorrem em <p<u>e</u>lo, p<u>e</u>la, p<u>e</u>los, p<u>e</u>las>, em que <e> se realiza [e].

Tabela 73: Regras de transcrição para os grafemas <j, k, l, m> do PB.

#	Padrão gráfico de <j>	Fone	Exemplo
1	...<j>...	[Z]	júnior
#	**Padrão gráfico de <k>**	**Fone**	**Exemplo**
1	...<k>...	[k]	kelvin
#	**Padrão gráfico de <l>**	**Fone**	**Exemplo**
1	...<l><V>...	[l]	ala
2	...<l h>...	[L]	alho
3	...<l>...	[w]	vogal
#	**Padrão gráfico de <m>**	**Fone**	**Exemplo**
1	...<e, é, ê, i><m><sp><V>...	[J]	Alguém usou, Quem está?
2	...<m>...	[m]	mameluco
#	**Padrão gráfico de <n>**	**Fone**	**Exemplo**
1	...<n h>...	[J]	ganho
2	...<n>...	[n]	nata

Tabela 74: Regras de transcrição para o grafema <o> do PB.

#	Padrão gráfico de <e>	Fone	Exemplo
1	...<(S)o><l><C-h,Pont>...	[O]	sol, girassol, futebol
2	...<(US)o><l><C-h,Pont>...	[o]	soldadura, soltar
3	...<ou>...	[ow]	ouvir, couve, estou
4	...<(S)o><a><Pont,Ltr>...	[o]	Lisboa, pessoa
5	...<o><so><Pont,Ltr>...	[o]	saudoso, virtuoso
6	...<o><sa><Pont,Ltr>...	[O]	saudosa, virtuosa
7	...<o,ô (m,n)><C-h,Pont>...	[o˜]	compositor,gôndola, som
8	...<(S)o><m,n>...	[o˜]	soma, sono, ponho
9	...<o><r><Pont>...	[o][197]	compor, dor
10	...<o><z><Pont>...	[O][198]	voz, algoz, atroz
11	...<ô>...	[o]	vovô
12	...<ó>...	[O]	vovó
13	...<õ>...	[o˜]	corações
14	...<(W bgn)c><o><Pont>...	[o]	co-produção
15	...<a><o><C,Pont>...	[w]	ao, caos

[197] Excepções: <maior(es), menor(es), melhor(es), pior(es), suor(es)> → [O].
[198] Excepção: <arroz> → [o].

Tabela 74: Regras de transcrição para o grafema <o> do PB (continuação).

16	...<o><Ltr><S_V>...	[o]	opor
17	...<o><Pont>...	[u]	músico
18	...<o><s><Pont>...	[u]	carros
19	...<o>...	[o]	escopo

Tabela 75: Regras de transcrição para os grafemas <p, q, r> do PB.

#	Padrão gráfico de <p>	Fone	Exemplo
1	...<ph>...	[f]	Philipe
2	...<p>...	[p]	pato
#	**Padrão gráfico de <q>**	**Fone**	**Exemplo**
1	...<qu><i,e,o>...	[k]	quito, quente, quota
2	...<q><ü,ua>...	[k]	cinqüenta, quase
#	**Padrão gráfico de <r>**	**Fone**	**Exemplo**
1	...<rr>...	[R]	carro
2	...<r sp r>...	[R]	Um pomar rodeado de flores.
3	...<(W bgn)r>...	[R]	A rua foi interditada.
4	...<r><V>...	[r]	ratoeira
5	...<r><sp><V,h>...	[r]	Falta acertar apenas uma.
6	...<r><sp><C_UV>...	[X]	Pecar pelo meio do caminho.
7	...<r><sp><C_VO>...	[R]	Injetar grãos de arroz.
8	...<r><C_UV>...	[X]	perco
9	...<r><C_VO>...	[R]	carga
10	...<r>...	[X]	Ela irá se lascar.

Tabela 76: Regras de transcrição para o grafema <s> do PB.

#	Padrão gráfico de <s>	Fone	Exemplo
1	...<tr(a,â)n><s><V>...	[z]	transação, trânsito
2	...<ob><s><équio>...	[z]	obséquio
3	...<ã><s><Pont>...	[j˜ s]	fãs, maçãs
4	...<S_V><s><Pont>...	[j s]	mas, gás, atrás, Goiás
5	...<sh>...	[S]	shiatsu
6	...<(W bgn)s>...	[s]	O sapato está lustrado.
7	...<V><s><V>...	[z]	asa
8	...<s><C_VO>...	[z]	transgredir

Tabela 76: Regras de transcrição para o grafema <s> do PB (continuação).

9	...<ss>...	[s]	a<u>ss</u>ar
10	...<sc><e,i>...	[s]	cre<u>sc</u>er
11	...<sç>...	[s]	cre<u>sç</u>am
12	...<s sp j>...	[Z]	Ele<u>s</u> jogaram bola.
13	...<s><sp><V,C v,h>...	[z]	O<u>s</u> aros são cromados.
14	...<s>...	[s]	Ele<u>s</u> reberam o prêmio.

Tabela 77: Regras de transcrição para os grafemas <t, u, v> do PB.

#	Padrão gráfico de <t>	Fone	Exemplo
1	...<th><Pont>...	[tS]	Ru<u>th</u>
2	...<th>...	[t]	Ar<u>th</u>ur
3	...<t><C-r,l>...	[tS]	algori<u>t</u>mo
4	...<t><i,[i] >...	[tS]	<u>t</u>ia, me<u>t</u>e
5	...<t><Pont>...	[tS]	Aquele se<u>t</u> foi difícil.
6	...<t>...	[t]	<u>t</u>a<u>t</u>o
#	**Padrão gráfico de <u>**	**Fone**	**Exemplo**
1	...<ü>...	[w]	ling<u>ü</u>ística
2	...<u (m, n)>...	[u˜]	ab<u>un</u>dante, ret<u>um</u>bante
3	...<u(m, n)><Pont>...	[u˜]	Ele come at<u>um</u>.
4	...<u><m, n>...	[u˜]	<u>u</u>ma, <u>u</u>nha
5	...<ú,u>...	[u]	ac<u>ú</u>stica
#	**Padrão gráfico de <v>**	**Fone**	**Exemplo**
1	...<v>...	[v]	<u>v</u>oando

Tabela 78: Regras de transcrição para os grafemas <w, x> do PB.

#	Padrão gráfico de <w>	Fone	Exemplo
1	...<w>...	[w]	<u>w</u>att
#	**Padrão gráfico de <x>**	**Fone**	**Exemplo**
1	…<x>…	[k s]	o<u>x</u>ítono, o<u>x</u>ítona, o<u>x</u>ítonos, o<u>x</u>ítonas, o<u>x</u>idar, í<u>x</u>ia, o<u>x</u>idação,comple<u>x</u>o, ane<u>x</u>ar, o<u>x</u>igênio, o<u>x</u>iúro, o<u>x</u>alato, ú<u>x</u>er, u<u>x</u>oricida, a<u>x</u>ila, a<u>x</u>iologia, tá<u>x</u>i, sinta<u>x</u>e
2	…<x> …	[k z]	i<u>x</u>ofagia, i<u>x</u>omielite, i<u>x</u>olite, i<u>x</u>ômetro, i<u>x</u>ora, i<u>x</u>oscopia, o<u>x</u>-acético

Tabela 79: Regras de transcrição para os grafemas < y, z> do PB.

#	Padrão gráfico de <y>	Fone	Exemplo
1	...<y><C>...	[i]	Yguacu
2	...<y>...	[j]	Yanomami
#	**Padrão gráfico de <z>**	**Fone**	**Exemplo**
1	...<z><sp><C_UV>...	[s]	Ferraz furou o ferro.
2	...<z><sp><C v,V,h>...	[z]	Faz anos que não o vemos; Ferraz gosta de pão; Faz horas que o vejo.
3	...<S_V><z><Pont>...	[j s]	faz
4	...<z><Pont>...	[s]	O que José fez?
5	...<z>...	[z]	zumbido

Os algoritmos acima representados foram testados com um extracto de 5465 palavras e 14920 fones do *corpus* Cetem-Folha. As transcrições fonéticas originadas pelo algoritmo foram manualmente verificadas, tendo-se obtido 97,44% de fones correctamente transcritos. Na Tabela 80, podem ver-se os resultados do teste efectuado. Tal como acontecia no PE, também no PB grande parte dos erros ocorre em palavras estrangeiras, em realizações de <e> e <o> comuns em homógrafos e em alternâncias vocálicas da vogal do radical verbal, em problemas de identificação de ditongos e de nasalizações vocálicas, sendo esta a causa de maiores erros, 0,84%. Estes resultados podem ser fortemente melhorados com a integração dos módulos de pré-processamento e desambiguação de homógrafos apresentados antes, cuja adaptação é praticamente directa, e com a implementação do leitor de estrangeirismos, após uma adaptação mais estudada.

Tabela 80: Erros resultantes do teste do transcritor grafema-fone com frases em PB.

Tipo de erro	# erros	% erros
Estrangeirismos	95	0,64
[e] ou [E]	55	0,37
[o] ou [O]	49	0,31
Ditongos	55	0,37
Acrónimos	5	0,03
Nasalização	125	0,84
Total	**384**	**2,56**

Do conjunto total de regras de conversão grafema-fone que descrevem o português europeu, 70,5% aplicam-se totalmente ao PB, 21,9% apresentam igual padrão de

input mas diferente padrão de output. Estes dados representam um total de 92,5% de compatibilidade entre as regras que descrevem a conversão grafema-fonema do PE e do PB.

5.6. Aplicações do sistema ao galego

5.6.1. Divisão silábica e marcação da sílaba tónica – testes e resultados

Os algoritmos propostos para o PE em 5.1. e 2.2. foram testados com textos aleatoriamente extraídos do CORGA[199] (Corpus de Referencia do Galego Actual). Este *corpus* é composto por vários tipos de textos extraídos de diversas fontes, desde jornais, livros, revistas e transcrições de corpora orais, o que permite abranger vários tipos de texto e várias áreas de conhecimento. O texto que serviu de teste ao nosso sistema contém 300 frases, 2627 palavras e 12250 caracteres sem espaços. Não foram feitas quaisquer adaptações do divisor silábico e do marcador de tónica ao Galego, excepto na lista de palavras átonas. Baseámo-nos para este teste nas Normas ortográficas e morfolóxicas do idioma galego, aprovadas pela Real Academia Galega (2003).

Da análise dos resultados, apresentados nas Tabelas 81 e 82, obtiveram-se taxas de acerto de 97,87% para o divisor silábico e de 98,52% para o marcador de tonicidade.

Tabela 81: Erros resultantes do teste do divisor silábico com frases em galego.

Tipo de erro	# erros	% erros
Por ausência de acento gráfico em galego	30	1,14
Não separação de hiatos	22	0,84
Separação de outros ditongos	4	0,15
Total	**56**	**2,13**

Nos dois testes, a maior parte dos erros decorre do facto de em galego, por influência do castelhano, se considerarem paroxítonas as palavras que terminam em ditongo /jo/ ou /ja/ (ex. <contrario, media, principio, Emilio, circunstancia>), palavras essas que no Português são consideradas proparoxítonas, e por isso, marcadas com acento gráfico (ex. <contrário, média, princípio, Emílio, circunstância>). Ora, o acento gráfico é essencial para a identificação da sílaba tónica nestes casos, e essa

[199] CORGA (Corpus de Referencia do Galego Actual) disponível em: http://corpus.cirp.es/corgaxml/ (11-12-2007). Estas frases foram gentilmente cedidas pelo Centro Ramón Piñeiro para a Investigación en Humanidades, uma vez que o que estava disponível ao público era o acesso ao corpus através de um sistema de consulta.

informação é importante também para que a separação silábica seja bem sucedida[200]. De todas as formas, a tendência do galego é pronunciar estas palavras como proparoxítonas, convertendo estes ditongos em hiatos (Freixeiro, 2006: 123).

Tabela 82: Erros resultantes do teste do marcador de tonicidade com frases em galego.

Tipo de erro	# erros	% erros
Por ausência de acento gráfico em galego	34	1,29
Estrangeirismos	2	0,08
Outros	3	0,11
Total	**39**	**1,48**

Assim, propomos as seguintes regras para a marcação da sílaba tónica (Tabela 83):

Tabela 83: Regras para a marcação da sílaba tónica em galego.

1	Se ^(0) = {a,o} e ^(1) = {i} e ^(2) é consoante e ^(3) = V → T = 3	contrario, media, principio
2	Se ^(0) = {s} e ^(1) = {a,o} e ^(2) = {i} e ^(3) é consoante e ^(4) = V → T = 4	contrarios, medias, principios
3	Se ^(0) = {a,o} e ^(1) = {i} e ^(2) é consoante e ^(3) = {m,n} e ^(4) = V → T = 4	circunstancia
4	Se ^(0) = = {s} e ^(1) = {a,o} e ^(2) = {i} e ^(3) é consoante e ^(4) = {m,n} e ^(5) = V → T = 5	circunstancias

As regras 19 e 21 da divisão silábica teriam de considerar esta marcação da sílaba tónica previamente. Esta simples adaptação representaria uma melhoria de desempenho imediata de 1,14% para o divisor silábico e de 1,29 para o marcador de tonicidade, elevando a taxa de acerto para 99,01 e 99,81% respectivamente para cada módulo, taxas muito próximas das obtidas no PE e PB.

Os restantes erros ao nível da divisão silábica decorrem, tal como em PE e PB, da não identificação de hiatos, que são interpretados como ditongos, em <adoecido> (a_doe_ci1_do) ou <bágoas> (bá1_goas). O terceiro tipo de erros tem a ver com a separação errada de ditongos em <ceo> (ce1_o) ou <traes> (tra1_es).

Quanto ao marcador de tonicidade, o segundo tipo ocorre em estrangeirismos, que tal como em PE e PB, são responsáveis por muitos erros (ex. <Madrid> ma1_drid) e em outros casos irregulares, como <mili> (mi_li1) e <seguimos> (se_gu1i_mos).

Estes testes e resultados não só demonstram a grande compatibilidade entre as estruturas fonológicas do Galego e do Português, como também a flexibilidade dos algoritmos propostos nesta dissertação.

[200] As regras afectadas nestes contextos e que não são activadas por falta de acento gráfico são as regras 19 e 21 da separação silábica e a regra 2 da marcação de tonicidade.

5.6.2. Transcritor grafema-fone – adaptação, testes e resultados

Para o galego, o sistema de síntese bilingue (para galego e castelhano) por concatenação, o Cotovía, desenvolvido pela Universidade de Vigo, parece ser, até este momento, o único sistema publicado. No entanto, o conjunto de regras de conversão grafema-fonema para o galego ainda não é conhecido.

Existe uma larga tradição de estudos de Fonética e Fonologia do galego com resultados experimentais reconhecidos (Freixeiro, 1998; Alvarez, 1991; Regueira, 1997). No entanto, no que respeita à aplicação deste conhecimento à Síntese da Fala para o galego, ainda existe muito trabalho por fazer, sobretudo a nível do módulo de pré-processamento e análise fonética.

A proximidade fonético-fonológica que o galego partilha com o português serviu-nos de motivação para experimentar a mesma metodologia.

Além disso, justificamos a nossa abordagem por regras linguísticas baseando-nos nas seguintes premissas: o galego, à semelhança do português, tem uma razoável regularidade fonológica; o galego, ao contrário do português, apresenta maior correspondência entre ortografia e fonética, devido a uma fixação ortográfica tardia, por motivos histórico-políticos, e devido ao ainda actual processo de normalização linguística; uma abordagem por regras linguísticas é mais económica em termos de recursos computacionais do que uma abordagem por dicionários; uma abordagem por regras linguísticas permite ler sempre uma palavra nova; um bom conjunto de regras baseado numa descrição fonológica é capaz de resolver praticamente qualquer problema de trascrição; finalmente, a nossa abordagem já foi aplicada ao português europeu e ao português do Brasil com uma taxa de acerto de 99,11% (Braga *et al.*, 2007b) e de 97,44% (Silva *et al.*, 2006) respectivamente, tendo a maior parte das regras sido adaptadas ao galego.

Ao longo da construção do nosso algoritmo, foram considerados estudos recentes em Fonética e Fonologia do galego, tendo sido incluídos alguns fenómenos de sandhi.

Esta adaptação ao galego foi já apresentada e publicada em:

- Braga, D, & Coelho, L. 2006. "Letter-to-sound conversion for Galician TTS systems", in *IV Jornadas en Tecnologías del Habla*, Zaragoza, Espanha. pp. 171-176.
- Braga, D.; Freixeiro, X. 2006. "Algoritmos de Conversão Grafema-Fone em Galego para Sistemas de Conversão Texto-Fala", in *VIII Congreso Internacional de Estudos Galegos,* Salvador, Bahía, Brasil, 12-15 de Setembro de 2006 (no prelo).

Os mesmos símbolos e convenções de anotação apresentados na Tabela 51 foram usados para a construção do algoritmo de transcrição grafema-fone para o galego.

Alguns casos foram previamente definidos, como vogais, consoantes, consoantes vozeadas e não vozeadas, vogais/sílabas tónicas e átonas e pronomes demonstrativos.

Para a anotação fonológica, apresentamos, na Tabela 84, uma proposta de alfabeto SAMPA para galego, baseada no conjunto de símbolos que compõem os alfabetos SAMPA para o português e para o espanhol, com algumas extensões: [l*] que representa a consoante lateral alveolar em situação implosiva presente em <sal>; o diacrítico /~/ para representar a nasalização provocada pela vizinhança de consoantes

nasais presente em <xente>; [z] e [T*] representando os alofones vozeados correspondentes às fricativas surdas /s/ e /T/ respectivamente.

Tabela 84: Proposta de SAMPA para galego.

classe de fonema	símbolo	exemplo
vogais	**/a/**	casa, cama, cal, acto
	/i/	río, fio, ir, digo
	/u/	xunta, último
	/E/	ferro, almacén, mel, exame,
	/e/	rochedo, ver, vexo
	/O/	voz, caracol
	/o/	curioso, don
semi-vogais	**/j/**	queixo, loita, aguɨa, fiestra
	/w/	louvar, mouro, frecuente
consoantes oclusivas	**/p/**	pé, campo
	/t/	trato
	/k/	carpa, aqui, kantiano
	/b/	bebo,
	/d/	caldo, morder
	/g/	gato, figo
consoantes africadas	**/tS/**	chamar, achar
consoantes fricativas	**/f/**	faro, feliz, café
	/T/	fácil, zarzallo, macio, luz
	/s/[201]	selo, cousa, cuspir
	/S/	xa, xente, muxica
consoantes nasais	**/m/**	memoria, campo
	/n/	nada, onte
	/J/	braña
	/N/	unha, algunha
consoantes líquidas	**/l/**	lúa, álxebra, colgar
	/L/	valla, muller
	/r/	coro, carta
	/rr/	rato, melro, aforrar

A nossa proposta de SAMPA para galego é de base fonológica e não fonética[202], apesar das extensões de tipo alofónico. A descrição fonológica revelou-se mais

[201] De facto, apesar de este símbolo ser comum aos SAMPAS do português e do espanhol, os dois fonemas têm realizações diferentes, sendo ápico-alveolar no caso do Espanhol e prédorso-velar no caso do português. O fonema galego aproxima-se mais da articulação espanhola.

optimizada, dada a grande variação alofónica presente no galego standard (Freixeiro, 2006: 188). Além disso, as variações de fonemas dependentes do contexto são parâmetros automaticamente extraídos aquando da selecção de unidades feita pelo treino da base de dados que alimenta o sistema de síntese.

Nas Tabelas seguintes numeradas de 85 a 92, apresentamos o conjunto de regras de transcrição fonológica para cada grafema presente na actual ortografia do galego, seguido de exemplos ilustrativos das respectivas ocorrências.

Tabela 85: Regras de transcrição para os grafemas <a, b, c, d> do galego.

#	Padrão gráfico de <a>	Fonema	Exemplo
1	... < a, á, à >...	/a/	casa, cama, cal, pau, acto, cá
#	**Padrão gráfico de <b>**	**Fonema**	**Exemplo**
1	... <b>...	/b/	ben, pobre
#	**Padrão gráfico de <c>**	**Fonema**	**Exemplo**
1	...<ch>...	/tS/	chuvia, achar, che
2	... < c > < e, i >...	/T/	certo, cinco
3	... <c>...	/k/	cando, caeu, actos, dialectal
#	**Padrão gráfico de <d>**	**Fonema**	**Exemplo**
1	... <d>...	/d/	despois, caldo, vida

Tabela 86: Regras de transcrição para o grafema <e> do galego.

#	Padrão gráfico de <e>	Fonema	Exemplo
1	... <SP, Pont> <e> <SP>...	/E/	ti e eu
2	...<(Prn_D), (S) e>...	/e/	este, ese
3	...<(W_bgn) e><x>...	/E/	excelente, exame
4	...<(S) e><C>< i,u><V/ {i,u}><Pont>...	/E/	seria, recua,
5	...<(S) e><C><C/ {m,n}>< i,u><V/i,u><Pont>...	/E/	modestia, controversia
6	...<e,>< n><C, Pont>...	/e/	meten, cólleno
7	...<é[203]>< n><Pont>...	/E/	tamén, vén
8	... <e> <n> < za, cio, cia><s, Pont>...	/E/	ciencia, pertenza
9	... < é > <s><Pont>...	/e/	lugués

[202] Em Losada Soto (2004), apresenta-se uma proposta muito completa de um alfabeto SAMPA para galego, de base fonológica, mas contemplando a vasta alofonia existente no galego falado.

[203] Também as palavras monossilábicas não acentudas graficamente <ten>, <quen> têm vogal aberta /E/.

Tabela 86: Regras de transcrição para o grafema <e> do galego (continuação).

10	... < e > <s, z> <a><s, Pont.>...	/e/	lugu<u>e</u>sa, avar<u>e</u>za
11	... < e > <do><s, Pont.>...	/e/	pen<u>e</u>do,
12	...<e><m,n><e><Pont>..	/E/	x<u>e</u>ne, l<u>e</u>me
13	...<(S) e><r><Pont>...	/e/	v<u>e</u>r, sab<u>e</u>r[204]
14	...<(S) e><u, o><Pont>....	/e/	s<u>e</u>o, fr<u>e</u>o, tem<u>e</u>u
15	...< (S) e ><la><Pont>...	/E/	v<u>e</u>la, cad<u>e</u>la, aqu<u>e</u>la
16	...<e><l><C, Pont>...	/E/	pap<u>e</u>l, p<u>e</u>l, f<u>e</u>lpa
17	...<i><e>...	/E/	fi<u>e</u>stra
18	...<(S) e><i>...	/e/	ch<u>e</u>ira, qu<u>e</u>ira
19	... < é >...	/E/	l<u>é</u>xico, vix<u>é</u>simo
20	...<é, e><(ct, cn,)>...	/E/	dial<u>e</u>cto, t<u>é</u>cnica
21	...<(US) e><o, a>...	/e/	ald<u>e</u>a, r<u>e</u>al, f<u>e</u>o
22	...<(S) e><ll, ñ, ch, x>...	/e/	v<u>e</u>xo, p<u>e</u>cho
23	... < (US) e >...	/e/	d<u>e</u>spois, <u>e</u>spañol
24	... < (US) e ><Ltr, Pont>...	/e/	chocolat<u>e</u>

Tabela 87: Regras de transcrição para os grafemas <f, g, h, i> do galego.

#	Padrão gráfico de <f>	Fonema	Exemplo
1	... < f >...	/f/	<u>f</u>ai, caf<u>é</u>, <u>f</u>lor, naí<u>f</u>
#	**Padrão gráfico de <g>**	**Fonema**	**Exemplo**
1	... < g u > < e, i >...	/g/	<u>gu</u>erra, <u>gu</u>indastre
2	... < g >...	/g/	fi<u>g</u>o, <u>G</u>aliza, fun<u>g</u>o
#	**Padrão gráfico de <h>**	**Fonema**	**Exemplo**
1	... < h >...	/ /	<u>h</u>oxe, <u>h</u>aber
#	**Padrão gráfico de <i>**	**Fonema**	**Exemplo**
1	...<i><ño, ña>...	/i~/	padr<u>iñ</u>o, morr<u>iñ</u>a
2	... <V/{i, u}> <i>...	/j/	cai<u>x</u>a, papei<u>s</u>, loi<u>t</u>a
3	...<i> <V/{i, u}>...	/j/	fi<u>e</u>stra, idi<u>o</u>ta, canción
4	... < i, í >...	/i/	r<u>í</u>o, f<u>í</u>o, <u>i</u>r, d<u>i</u>go

[204] Excepção: <muller> →/muLEr/, <culler> → /kuLEr/.

Tabela 88: Regras de transcrição para os grafemas <l, m, n, ñ > do galego.

#	Padrão gráfico de <l>	Fonema	Exemplo
1	... < l > <C/h, Pont>...	/l*/	calma, último, alzar, colgar
2	... < ll >...	/L/[205]	muller, mollado, vello
3	...<l>...	/l/	lúa, valado, vela
#	**Padrão gráfico de <m>**	**Fonema**	**Exemplo**
1	...<V><m><C>...	/~m/	cumprido, importante, tempo, ambos
2	... < m >...	/m/	memoria, campo
#	**Padrão gráfico de <n>**	**Fonema**	**Exemplo**
1	... <u><nh><a>...	/N~/	unha, algunha
2	...<V><n><C, Pont>...	/~n/	onte, antes, don,
3	... <n>...	/n/	nai, neno, onte, once, onda
#	**Padrão gráfico de <ñ>**	**Fonema**	**Exemplo**
1	...<ñ>...	/~J/	braña, niño, mariñeiro

Tabela 89: Regras de transcrição para o grafema <o> do galego.

#	Padrão gráfico de <o>	Fonema	Exemplo
1	1	....<ó>...	/O/[206]
2	2	…< (W_bgn) h> <o> <r, s, t>…	/O/[207]
3	3	...<o, ó><n>><es, C, Pont>...	[o]
4	4	…<(S) o><r><es, Pont>…	/o/
5	5	…<o><z>< Pont>…	/O/
6	6	…<o><ces>< Pont>…	/O/
7	7	…<o><so><s, Pont>…	/o/
8	8	…<o><sa> <s, Pont>…	/O/

[205] Este fonema lateral palatal sonoro vem sendo substituído pelo fricativo palatal sonoro /ʝ/ devido à influência do castelhano sobretudo nos meios urbanos: "Este fenómeno denomínase deslateralización e pode xa constituír un paso irreversibel no noso sistema fonolóxico, se é que non se lle pon freo á influencia fonética do castelán, pois é o seu triúnfo case que definitivo nesa lingua o que provoca a extensión polo territorio de fala galega, afástandonos máis unha vez do portugués, onde /ʎ/ se conserva" (Freixeiro, 2006: 180). Este fenómeno de deslateralização não é aceite no galego standard.

[206] Excepção a esta regra: formas do Infinitivo de <pór> /o/ e derivados.

[207] Excepção a esta regra: <hoxe> /o/.

Tabela 89: Regras de transcrição para o grafema <o> do galego (continuação).

9	9	…<o><l, la><C, Pont>…	/O/
10	10	…<o><i><s><Pont>…	/O/
11	11	.. ..<(S) o><C>< i ><V/ {i,u}><Pont>...	/O/
12	12	...<o><u>...	/o/
13	13	...<o><i>...	/o/
14	14	…<Ltr><o><l><C/h>…	/o/
15	15	...<(S) o><a>...	/o/
16	16	...<o><(pt, bt, bx, ct, cl, gn)>...	/O/
17	17	... <(US) o > <s, Ltr, Pont, SP>...	/o/
18	19	... <(US) o >...	/o/

Tabela 90: Regras de transcrição para os grafemas <p, q, r, s, t> do galego.

#	Padrão gráfico de <p>	Fonema	Exemplo
1	... < p >...	/p/	pato, polo, carpa
#	**Padrão gráfico de <q>**	**Fonema**	**Exemplo**
1	...<qu><e,i>...	/k/	esquina, quen, queixo
2	... < q >...	/k/	quórum
#	**Padrão gráfico de <r>**	**Fonema**	**Exemplo**
1	... < r r > ...	/rr/	aforrar, curro
2	...<(W_bgn) r>...	/rr/	rúa, ría
3	... <l,n><r>…	/rr/	honra, melro
4	... < r >...	/r/	mar, cara, fraco
#	**Padrão gráfico de <s>**	**Fonema**	**Exemplo**
1	…<s><C_VO>…	/z/[208]	prosma, faslo
2	…<s><SP><C_VO>…	/z/	estás desperto
3	...<s>...	/s/	selo, cousa, cuspir
#	**Padrão gráfico de <t>**	**Fonema**	**Exemplo**
1	… <t>…	/t/	corte, tamén

[208] Extensão SAMPA para representar a fricativa alveolar sonora.

Tabela 91: Regras de transcrição para os grafemas <u, v, x> do galego.

#	Padrão gráfico de <u>	Fonema	Exemplo
1	... < ü > ...	/w/	lingüista, pingüe
2	... <V/u>< u >...	/w/	cauto, meu, pouco, fuxiu
3	...<u><V/u>	/w/	minguar, sueco, tenue, frecuente
4	... < ú, u >...	/u/	túa, rúa, pulo, miúdo
#	**Padrão gráfico de <v>**	**Fonema**	**Exemplo**
1	... < v >...	/b/	avogado, volveu, vostede
#	**Padrão gráfico de <x>**	**Fonema**	**Exemplo**
1	Lista de excepções	/ks/	axila
2	...<(W_bgn) ex, exo, extra, taxi, xeno, xilo>...	/ks/	extra, exame, excedente, xenofobia, xilófono
3	...< (W_bgn) ex><V><x>...	/ks/	exaxerar, exixir, exexese, exixencia
4	...< x>...	/S/	xamón, xente, obxecto, reloxio
#	**Padrão gráfico de <z>**	**Fonema**	**Exemplo**
1	... < z SP> <C_VO>...	/T*/[209]	cruz verde, cruz branca
2	... <z>...	/T/	zunir, prezo, luz

Tabela 92: Regras de transcrição para os grafemas <j, k, y, w> do galego.

#	Padrão gráfico de <j>	Fonema	Exemplo
1	... <j>...	/dZ/	jeep, jazz, judo
#	**Padrão gráfico de <k>**	**Fonema**	**Exemplo**
1	... < k>...	/k/	karate, kamikaze
#	**Padrão gráfico de <y>**	**Fonema**	**Exemplo**
1	….<(W_bgn) y>…	/jj/	yin, yang
2	...<y>...	/i/	baby, bodyboard
#	**Padrão gráfico de <w>**	**Fonema**	**Exemplo**
1	…<w>…	/w/	walkie-talkie, whisky

[209] Extensão SAMPA para representar a fricativa interdental sonora.

O alfabeto do galego é composto por 23 grafemas e 6 dígrafos (<ch>, <gu>, <ll>, <nh>, <qu>, <rr>), tendo os útimos sido incluídos nos padrões gráficos dos grafemas simples.

Foram considerados todos os padrões gráficos documentados na literatura, incluindo os grafemas <j>, <k>, <y>, <w> (cf. Tabela 92) que ocorrem nos estrangeirismos.

Também foram tidos em consideração os ditongos crescentes e decrescentes, que foram incluídos nas regras de transcrição dos grafemas <i> e <u>. Grande parte das regras apresentadas estão descritas na literatura (Freixeiro, 2006) de referência para o galego standard. Cada saída fonológica foi verificada por um linguista de galego, assegurando assim o rigor da transcrição de todos os padrões gráficos.

Ao conceber estas regras, tentámos cingir-nos o mais possível às sequências apresentadas pelos padrões gráficos, procurando assim reduzir a dependência do transcritor grafema-fonema de outros módulos presentes no conversor grafema-fone, como o marcador de sílabas tónicas e o divisor silábico. As principais excepções às regras descritas neste trabalho são também apresentadas em notas de rodapé.

As transcrições dos 27 grafemas e dos 6 dígrafos mencionados perfazem um total de 93 regras.

A maior parte dos problemas com que nos defrontámos ao construir o conversor grafema-fonema para galego são os mesmos que encontrámos no português e prendem-se com a problemática das alternâncias vocálicas entre /e/ e /E/, por um lado, e /o/ e /O/, por outro.

Estas alternâncias vocálicas podem ocorrer nos seguintes casos:

- no interior da flexão verbal (ex. <meto> /e/ vs <metes> /E/);
- em pares de homógrafos heterófonos quer apresentem categorias gramaticais diferentes (por exemplo, nome vs verbo: <comezo> /e/ vs <comezo> /E/; <acordo> /o/ vs <acordo> /O/), quer possuam a mesma categoria gramatical (por exemplo, nome vs nome: <besta> /e/ vs <besta> /E/);
- em palavras afectadas por metafonia de <o>, fenómeno fonético amplamente difundido no galego e no português medievais, que consiste numa influência assimilatória regressiva operada pela vogal átona final fechada sobre a vogal tónica aberta. A vogal final provoca o fechamento da vogal tónica: /E/>/e/ (ex. <medo>) e /O/>/o/ (ex. <novo>).

A própria etimologia da palavra é também responsável por encontrarmos um determinado timbre vocálico na vogal tónica (por exemplo, a vogal tónica de <ferro> é /E/, porque deriva de FĔRRU- ao passo que de <negro> é /e/), porque deriva de NĬGRU-).

O grafema <x>, por exemplo, é um dos casos mais difíceis de transcrever por razões etimológicas. Segue-se uma lista de algumas excepções às regras apresentadas na Tabela 9 para o grafema <x> em que este se pronuncia /ks/: anglosaxón, aproximar, asfixia, axila, axilar, axioma, bórax, clímax, complexo, convexo, crucifixo, elixir, exacto, exame, exaxerar, exceder, excelente, excepto, exclamar, exilio, eximir, éxodo, exótico, explosión, expurgar, extenso, extensión, extensor,

extra, fax, galaxia, hexágono, laxo, léxico, maxilar, nexo, reflexionar, prefixo, reflexo, saxón, sexo, sexual, sílex, sintaxe, sufixo, taxi, texto, textual, tórax, tóxico.

O algoritmo proposto foi testado com textos aleatoriamente seleccionados do CORGA (Corpus de Referencia do Galego Actual). O texto que serviu de teste ao nosso sistema contém um número total de 11245 caracteres distribuídos por 2387 palavras. Os fonemas originados pelo nosso sistema foram manualmente corrigidos e 98,50% foram correctamente convertidos. Os erros identificados e classificados estão apresentados em frequência absoluta e em percentagem relativa em relação ao número total de grafemas na Tabela 93.

De um total de 1,50% de erros (168 grafemas apresentam erros de transcrição), a maior parte ocorre, tal como acontece emn português, ao nível das realizações dos grafemas <e> e <o> em situação pré-tónica e tónica (0,93%), devido aos factores atrás mencionados que têm a ver com a morfologia verbal, metafonia e etimologia. A homografia heterofónica, presente em casos como <de acordo> /o/ vs <eu acordo> /O/, constitui também um problema representado em 0,09% de erros.

Tabela 93: Resultados do transcritor grafema-fonema para galego.

Tipo de erro	# occur.	% occur.
Erros na desambiguação de homógrafos com <e>	6	0,05
Erros na desambiguação de homógrafos com <o>	5	0,04
Erros no <e> por morfologia verbal	17	0,15
Erros no <o> por morfologia verbal	14	0,12
Erros no <e> por etimologia	28	0,24
Erros no <o> por etimologia	35	0,31
Erros na transcrição de <x>	12	0,10
Excepções não implementadas e outros problemas	51	0,45
Total	**168**	**1,50**

Questões como esta podem ser resolvidas com auxílio de um analisador morfológico, visto que a maior parte dos homógrafos pode ser desambiguado a partir da distinção gramatical entre nome/verbo. Também no caso de alternâncias vocálicas encontradas ao longo das conjugações verbais (ex. <eu bebo> /e/, <ti bebes> /E/, <el bebe> /E/), a informação linguística, de ordem morfológica e fonológica, poderá ajudar a reduzir a taxa de erro. São ainda de salientar os erros provenientes de não implementação de excepções, ou de outros casos, como a leitura de acrónimos, com

um total de 0,45%. O grafema <x>[210], à semelhança do que acontece no português, é também um caso em que se verificam erros de transcrição, na ordem dos 0,12%.

Os resultados da Tabela 93, quando comparados com os sistemas de conversão para português europeu (Braga *et al.*, 2007b) e português do Brasil (Silva *et al.*, 2006), permitem confirmar a boa performance do nosso sistema e o sucesso da abordagem por regras linguísticas.

Da análise dos resultados obtidos, pensamos que uma análise morfológica poderá melhorar o desempenho do nosso sistema ao nível da desambiguação de homógrafos. A informação fonológica do vocalismo tónico e pré-tónico ao longo da conjugação está prevista em desenvolvimentos futuros. Estes e outros problemas são comuns ao galego, ao português europeu e ao português do Brasil. Embora, para o galego, não tenhamos considerado os estrangeirismos, esta é também uma questão complexa nas três variedades cujo estudo está também em curso. Apesar de já terem sido considerados alguns fenómenos de sandhi, nomeadamente ao nível da fonologia sintáctica das consoantes, prevemos completar o nosso algoritmo com os fenómenos ao nível da vizinhança fonológica entre vogais.

Do conjunto total de regras de conversão grafema-fone que descrevem o português europeu, 80,2% aplicam-se totalmente ao caso do galego, 19,1% apresentam igual padrão de input mas diferente padrão de output. Estes dados representam um total de 95,3% de compatibilidade entre as regras que descrevem a conversão grafema-fonema do galego e do português.

5.7. Síntese do capítulo 5

Como síntese deste capítulo, destacamos os seguintes tópicos:

- A questão da conversão grafema-fone é um assunto que está longe de estar resolvido, como se poderá verificar pelo grande número de publicações sobre o tema na comunidade científica lusófona;
- Incluímos no conversor grafema-fone os seguintes sub-módulos: o divisor silábico, o marcador de sílaba tónica e o transcritor fonético;
- Propusemos algoritmos baseados em regras linguísticas para resolução da problemática da conversão grafema-fone/ma em PE, PB e galego;
- O algoritmo de marcação de tonicidade, constituído por 31 regras, representa uma das principais inovações deste trabalho, com resultados muito interessantes;
- Outra aspecto inovador deste trabalho é a apresentação exaustiva e detalhada do conjunto de regras linguísticas para a transcrição grafema-fone/ma em PE, do PB e do galego;

[210] Em galego, <g> etimológico seguido de <e, i> e <j> etimológico convergiram para a mesma forma gráfica <x>.

- Testes efectuados a cada módulo com textos reais em PE revelaram as seguintes taxas de acerto: 99,06% para o divisor silábico, 99,54 % para o marcador de sílaba tónica e 99,11% para o transcritor grafema-fone;
- Quando testados com textos reais em PB, as taxas de acerto do divisor silábico e do marcador de tónica, usados sem quaisquer adaptações, foram de 99,20% e 99,60% respectivamente;
- Do conjunto total de regras de conversão grafema-fone que descrevem o PE, 70,5% aplicam-se totalmente ao PB e 21,9% apresentam igual padrão de input mas diferente padrão de output;
- Os testes do transcritor grafema-fone para PB revelaram uma taxa de acerto de 97,44%;
- As taxas de acerto do divisor silábico e do marcador de tónica sem quaisquer adaptações quando testados com textos reais em galego foram de 97,87% e 98,52 respectivamente;
- Do conjunto total de regras de conversão grafema-fone que descrevem o PE, 80,2% aplicam-se totalmente ao caso do galego e 19,1% apresentam igual padrão de input mas diferente padrão de output;
- Os testes do transcritor grafema-fone para galego revelaram uma taxa de acerto de 98,50%;
- Estes testes e resultados não só demonstram a grande compatibilidade entre as estruturas fonológicas do PE, PB e galego, como também a flexibilidade dos algoritmos propostos.

Capítulo 6

Integração do sistema no motor de síntese

Neste capítulo, descreve-se o processo de integração dos vários módulos de normalização de texto e análise linguística, apresentados ao longo desta dissertação, num sistema de conversão texto-fala, com o objectivo de demonstrar a sua usabilidade e aplicabilidade.[211]

6.1. Construção e gravação da voice font

A *voice font* ou base de dados de fala é extremamente importante para a qualidade final do sistema de síntese da fala. São inúmeros os factores a ter em conta para a construção de qualquer *voice font*, a saber:

- A selecção do locutor (idade, sexo, dialecto, nível de escolaridade, local onde fez a escolaridade e onde viveu a maior parte da vida, experiência de locução);
- Qualidade subjectiva da voz (agradabilidade, atitude, sensualidade, expressividade, inteligibilidade, aplicabilidade, débito, idade subjectiva);
- Qualidade objectiva da voz (ao nível dos valores de frequência fundamental, intensidade, débito, *shimmer*[212], *jitter*[213]);

[211] Este trabalho contou com a colaboração de Ranniery Maia e Luís Coelho, responsáveis pelo desenvolvimento dos motores de síntese por HMMs em PB e em PE, e a quem dirigimos os nossos agradecimentos.

[212] São essencialmente variações de amplitude do sinal em períodos consecutivos e considerando o sinal estacionário. Há várias formas de calcular o *shimmer*. No Praat, software de análise de sinal bastante difundido (disponível em: http://www.fon.hum.uva.nl/praat/ - 03-01-2008), a forma de cálculo básica (Shimmer ddp) é: "average absolute difference between consecutive diferences between the amplitudes of consecutive periods."

[213] São essencialmente variações do período em períodos consecutivos e considerando o sinal estacionário. Há várias formas de calcular o *jitter*. No Praat, a forma de cálculo básica (Jitter

- Selecção dos *scripts* a gravar (tipos de texto, estilos, tipos de frases, variação prosódica e emotiva) e dimensão do *script* final em número de palavras;
- Características técnicas do estúdio e da gravação (equipamento, nível de insonorização da cabine de gravação, contacto visual da cabine para a régie, modelo de microfone, inserção ou não de filtros durante a gravação, frequência de amostragem da gravação, gravação ou não de canal para captação do sinal do EGG[214] e escolha do modelo de electroglotógrafo);
- Definição de especificidades ao nível da edição (validação do conteúdo linguístico, eliminação dos ruídos de diferente natureza que podem ser captados na gravação, como cliques, batidas no microfone, salivação, regravação deste tipo de erros, downsampling[215]).

De entre algumas publicações em que se descrevem os processos de construção de bases de dados para o português, destacamos Teixeira *et al.* (2001), sobre uma base de dados de cerca de 20 minutos etiquetada foneticamente para o PE e Cirigliano *et al.* (2005), em que se propõem 1000 frases foneticamente balanceadas para PB. Em Braga *et al.* (2007b e 2007e), apresenta-se o processo de selecção de locutores de voz para sistemas de síntese da fala em PB e PE e as características objectivas e subjectivas que permitem descrever uma voz de alta qualidade nestas duas variedades.

Pelos factores acima listados, pode verificar-se que a *voice font* é um recurso muito caro, que resulta de um processo demorado e que envolve vários intervenientes com tarefas muito especializadas (locutor, técnico auxiliar de gravação, técnicos de edição, linguistas para validação, engenheiros/linguistas para selecção de *scripts*).

A dimensão e características da *voice font* estão em grande parte dependentes da técnica de síntese utilizada. A técnica de síntese por concatenação é largamente aplicada e tem revelado uma qualidade competitiva, como se pode observar pelos resultados da avaliação feita a sistemas comerciais pela ASR News[216] e pelos resultados da avaliação a sistemas académicos divulgados nos Blizzard Challenges (Black & Tokuda, 2005; Bennet & Black, 2006; Fraser & King, 2007). No entanto, esta técnica exige bases de dados de fala muito extensas, com pelo menos 10000 frases, e consecutivamente muito dispendiosas.

A síntese por HMMs, que constitui já o novo paradigma das técnicas de síntese (Tokuda *et al.*, 1995; Tokuda, 2004), tem demonstrado ser capaz de produzir voz sintética com qualidade competitiva (Bennet & Black, 2006), recorrendo a bases de dados muito reduzidas (em Maia *et al.*, 2003 reportam-se 80 frases; evoluções desse trabalho, em Maia *et al.*, 2006 reportam 221 frases). A descrição e demonstrações desta técnica podem ser encontradas na página oficial do HTS[217] e no capítulo de Taylor (2007) sobre "HMM synthesis".

---

ddp) é: "average absolute difference between consecutive diferences between consecutive periods, divided by the average period."

[214] EGG – Electroglotograph (sinal que resulta da oclusão das cordas vocais).

[215] Conversão de uma frequência de amostragem em outra mais pequena.

[216] Resultados de "Text-to-Speech Accuracy Testing - 2005" e "Text-to-Speech Accuracy Testing – 2006" disponíveis em: http://www.asrnews.com/accuracy.htm (08-12-2007).

[217] Disponível em: http://hts.sp.nitech.ac.jp/ (08-12-2007).

Assim, tendo em vista a utilização desta técnica aplicada ao PB por Maia *et al.* (2003), Maia (2006) e Maia *et al.* (2006) e iniciada para o PE por Barros *et al.* (2005), construímos uma base de dados de fala com 1000 frases, extraídas aleatoriamente do *corpus* do Cetem-Público, através de um algoritmo de busca que selecciona frases entre 8 e 12 palavras, desenvolvido por Luís Coelho, em trabalho não publicado, baseado na proposta de Cirigliano *et al.* (2005). Estes *scripts* foram gravados por uma locutora de sexo feminino, com 28 anos, professora universitária e linguista, que viveu e estudou até ao nível universitário no Porto, Portugal, utilizando a variedade padrão do PE.

Foram eliminadas as frases com as seguintes características:

- Frases incompletas que por isso perderam o sentido;
- Frases com palavras de difícil leitura (certos estrangeirismos, siglas extensas ou nomes de marcas) que pudessem causar hesitações durante a locução.

A gravação teve lugar em estúdio profissional situado nas instalações da Universidade da Coruña, em Junho de 2006, tendo sido utilizado o seguinte equipamento: microfone de condensador Behringer B-2, mesa de misturas Behringer EuroRack MX602 A, gravador digital Grundig DCC 305, gravação feita a 44.1 KHz 16 bits de frequência de amostragem, transferência directa para computador através de porta SPDIF, edição com software Adobe Audition 1.5. Não foi captado sinal de EGG. A gravação foi realizada ao longo de cerca de 5 horas tendo sido obtidos 75 minutos e 41 segundos, incluindo silêncios, de tempo total de gravação após a edição, que foi também realizada no mesmo estúdio da Universidade da Coruña. A base de dados foi posteriormente etiquetada automaticamente, utilizando um reconhecedor de voz por HMMs descrito em Maia (2006) e corrigidas manualmente.

6.2. Integração do sistema com o motor de síntese por HMMs

A arquitectura e funcionamento do sintetizador de fala por HMMs aplicado ao PB é amplamente descrito em Maia (2006) e em Maia *et al.* (2006) e é fruto da colaboração estreita com o LPS – UFRJ, nomeadamente com Ranniery Maia e recentemente com Luís Coelho, em trabalho de doutoramento em curso sobre a aplicação da técnica de síntese por HMMs ao PE.

Não é nosso objectivo descrever o mecanismo de funcionamento do sintetizador por HMMs, que pode ser visto em Maia (2006), mas a Figura 49 permite ilustrar a sua arquitectura. Em linhas muito gerais, existe uma fase prévia de treino dos modelos de HMMs, que tem como inputs o registo acústico do sinal de fala e a informação linguística, fase essa que serve de base ao funcionamento do sistema. Após o treino, o texto que entra no sistema é processado pelo *front-end* que devolve etiquetas fonéticas que por sua vez são transformadas pelo módulo de selecção e concatenação de HMMs. Em seguida é aplicado o módulo de determinação de durações, seguido do módulo de geração de parâmetros acústicos que se divide em dois tipos: o módulo que

gera a fonte, ou o sinal de excitação ou os impulsos glotais, e o módulo que gera o filtro, ou seja, o módulo que simula as constrições ao longo do tracto vocal.

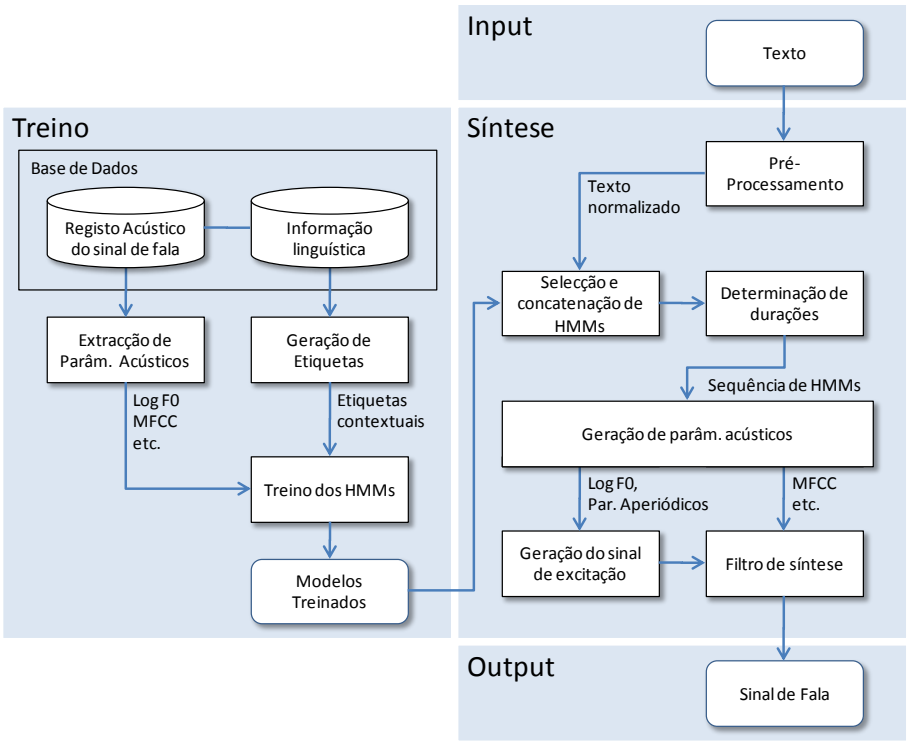

Figura 49: Arquitectura de um sistema de síntese por HMMs.

Mostraremos em seguida de que forma o sistema usa a informação decorrente dos módulos de *front-end* apresentados nesta dissertação.

O sistema de síntese por HMMs pode ser dividido em dois processos: uma parte em que o sistema é treinado e outra parte em que ocorre a síntese propriamente dita.

A primeira parte de treino consiste em três fases: 1) a extracção de parâmetros a partir da base de dados de fala; 2) a conversão da informação linguística das frases/enunciados da base de dados em etiquetas contextuais de HMMs; 3) o treino dos HMMs.

A segunda parte de síntese envolve as seguintes fases: 1) geração de etiquetas a partir da informação dos enunciados/frases; 2) selecção e concatenação por HMMs; 3) determinação de parâmetros; e 4) controlo da excitação e filtragem.

Para ambas as partes, de treino e de síntese, o sistema necessita de informações linguísticas dos enunciados/frases que compõem a base de dados, nomeadamente:

- Lista de fones da língua (alfabeto em SAMPA – vide Tabela 1);
- Divisão silábica fonética e marcação de tonicidade;

- Transcrição fonética da palavra e informação sobre a sua categoria morfossintáctica.

Na Figura 50, pode ver-se um exemplo da informação linguística utilizada pelo sistema que é gerada automaticamente a partir dos módulos de conversão grafema-fone (englobando os vários módulos integrados: o pré-processamento, leitor de estrangeirismos e desambiguador de homógrafos), divisor silábico e marcador de tonicidade apresentados ao longo deste trabalho. Como pode observar-se, o sistema gera uma tabela de etiquetas para cada frase da base de dados com informação de fone (phone), divisão silábica (syll), informação de sílaba tónica (representada por 1) ou átona (representada por 0) (stress), informação de fronteira de palavra (word) e informação de categoria morfossintáctica da palavra (class).

phone	syll	stress	word	class
sil				
E	E	1	ElA	function
l				
A	lA	0		
A	A	0	Ate~d@	content
t	te~	1		
e~				
d	d@	0		
@				
A	A	0	A	function
A	A	0	AkAd@miA	content
k	ka	0		
A				
d	d@	0		
@				
m	mi	1		
i				
A	A	0		
s	swe	1	swEkA	content
w				
E				
k	ka	0		
A				
sil				

Figura 50: Informação linguística gerada automaticamente para a frase "Ela atende a Academia Sueca."

A informação de classe de palavra (class) é obtida através do output do analisador morfológico utilizado pelo desambiguador de homógrafos nesta dissertação e apenas apresenta duas categorias: palavras função (function), constituídas pela biblioteca de classes fechadas (preposições, artigos, pronomes, conjunções, interjeições) e palavras conteúdo (content), constituída por nomes, verbos, adjectivos e advérbios. Em Maia (2006) ficou demonstrado que as informações de divisão silábica e marcação de tónica, por um lado, e de POS ou categoria morfossintáctica, por outro, incrementavam substancialmente a qualidade subjectiva da voz sintética. Mais uma vez fica demonstrada a importância deste tipo de informação linguística e a necessidade da existência de ferramentas que permitam prever, de forma automática, este tipo de informações fonológicas.

Além destas informações, o sistema usa ainda as seguintes informações linguísticas, designadas por "contextual factors" (Maia, 2006), destinadas à geração prosódica automática:

- Ao nível do fone:
  - segundo fone anterior, fone anterior, fone corrente, fone posterior, segundo fone posterior;
  - posição do fone corrente na sílaba corrente;

- Ao nível da sílaba:
  - informação sobre se a sílaba anterior, corrente e posterior é tónica ou não;
  - número de fones nas sílabas anterior, corrente e posterior;
  - posição da sílaba na palavra corrente;
  - número de sílabas tónicas na frase corrente antes e depois da sílaba corrente;
  - número de sílabas contando da tónica anterior à sílaba corrente;
  - número de sílabas, contando da corrente à sílaba tónica seguinte;

- Ao nível da palavra:
  - POS das palavras anterior, corrente e posterior;
  - número de sílabas das palavras anterior, corrente e posterior;
  - posição da palavra corrente na frase corrente;
  - número de palavras conteúdo nas frases anterior, corrente e posterior;
  - número de palavras contando da palavra conteúdo anterior à palavra corrente no enunciado;
  - número de palavras contando a partir da corrente até à palavra conteúdo seguinte no enunciado;

- Ao nível da frase:
  - número de sílabas, palavras nas frases anterior, corrente e posterior;
  - posição da frase corrente no enunciado corrente;

- Ao nível do enunciado
  - número de sílabas, palavras, frases.

Finalmente, o sistema formula questões acerca dos traços fonéticos inerentes aos 38 fones do português[218]. Segue a tabela classificativa das características articulatórias dos fones do PE (Tabela 94) (Mateus *et al.*, 1990: 328; 343).

[218] Exemplos de questões para PB: "Is current phone a voiced fricative?, Is pre-preceding phone voiced? Is succeeding phone an oral semi-vowel? Is post-succeeding phone a convex alveolar consonant?" (Maia *et al*., 2006).

Tabela 94: Classificação articulatória dos fones do PE.

SAMPA	Consoantes
p	consoante oclusiva bilabial
b	consoante vozeada oclusiva bilabial
t	consoante oclusiva alveolar
d	consoante vozeada oclusiva alveolar
k	consoante oclusiva velar
g	consoante vozeada oclusiva velar
f	consoante fricativa labiodental
v	consoante vozeada fricativa labiodental
s	consoante fricativa alveolar
z	consoante vozeada fricativa alveolar
S	consoante fricativa pós-alveolar
Z	consoante vozeada fricativa pós-alveolar
m	consoante vozeada soante nasal bilabial
n	consoante vozeada soante nasal alveolar
J	consoante vozeada soante nasal palatal
l	consoante vozeada soante lateral alveolar
L	consoante vozeada soante lateral palatal
r	consoante vozeada soante vibrante alveolar
R	consoante vozeada soante vibrante uvular
l*	consoante vozeada soante lateral velarizada alveolar
SAMPA	**Vogais e semivogais**
i	vogal soante vozeada anterior fechada
e	vogal soante vozeada anterior semi-fechada
E	vogal soante vozeada anterior semi-aberta
a	vogal soante vozeada central aberta
6	vogal soante vozeada central semi-aberta
O	vogal soante vozeada posterior semi-aberta
o	vogal soante vozeada posterior semi-fechada
u	vogal soante vozeada posterior fechada
@	vogal soante vozeada central fechada
w	glide soante vozeada posterior fechada
j	glide soante vozeada anterior fechada
j~	glide soante vozeada nasal anterior fechada
w~	glide soante vozeada nasal posterior fechada
i~	vogal soante vozeada nasal anterior fechada
e~	vogal soante vozeada nasal anterior semi-fechada
ɐ~	vogal soante vozeada nasal central semi-aberta
o~	vogal soante vozeada nasal posterior semi-fechada
u~	vogal soante vozeada nasal posterior fechada

6.3. Síntese do capítulo 6

Como resumo dos principais tópicos apresentados neste capítulo, destacamos:

- A construção de gravação da voice font é um dos aspectos mais importantes para a qualidade final da voz sintética, pelo que a sua recolha e construção deve ter em atenção muitos factores, nomeadamente: a selecção do locutor, a qualidade subjectiva e objectiva da voz, a selecção dos *scripts* a gravar, as características técnicas do estúdio e da gravação, a definição de especificidades ao nível da edição;
- A dimensão e características da *voice font* estão em grande parte dependentes da técnica de síntese utilizada; a técnica concatenativa, muito difundida e com qualidade muito competitiva, exige porém maiores bases de dados, enquanto a síntese por HMMs tem produzido voz sintética com qualidade não inferior recorrendo a bases de dados muito menores;
- Com vista a fazer a demonstração de integração dos algoritmos apresentados neste trabalho, construiu-se uma base de dados com 1000 frases repartidas ao longo de 75 minutos e 41 segundos, que foi automaticamente etiquetada e corrigida manualmente;
- A base de dados e os algoritmos de análise de texto apresentados neste trabalho e que constituem o *front-end* foram finalmente integrados no motor de síntese por HMMs, em dois momentos: uma fase para o treino e outra fase para a síntese.

Capítulo 7

Conclusões e trabalho futuro

> Uma língua é o lugar donde se vê o Mundo e em que se traçam os limites do nosso pensar e sentir. Da minha língua vê-se o mar. Da minha língua ouve-se o seu rumor, como da de outros se ouvirá o da floresta ou o silêncio do deserto. (Vergílio Ferreira)

Neste trabalho, descreveu-se o desenvolvimento dos módulos que constituem quer o pré-processamento ou normalização de texto (separação de frases e de palavras, expansão de abreviaturas, conversão de símbolos e caracteres especiais, conversão de siglas e acrónimos, leitura de numerais árabes cardinais e ordinais, leitura de números romanos, leitura de horas, datas, números com casas decimais, medidas e pontuação desportiva), quer a análise fonética (desambiguação de homógrafos, leitura de estrangeirismos, divisão silábica, marcação de sílaba tónica e transcrição grafema-fone) de um sintetizador de fala em PE e discutiram-se as suas aplicações ao PB e ao galego.

Os objectivos inicialmente propostos foram cumpridos:

- Descrição do estado da arte geral da área da síntese da fala no plano internacional e nacional e descrição do estado da arte específico para todos os módulos que compõem a normalização de texto e a análise fonética de um sintetizador de fala, aqui apresentados;
- Enquadramento do tema da presente dissertação na arquitectura de um sistema de síntese da fala;
- Proposição de regras de pré-processamento de texto, nomeadamente:
    - regras de separação de frases e de palavras para PE, aplicáveis também ao PB e ao galego;
    - regras de conversão de símbolos e expansão de abreviaturas para PE, sua implementação, teste e adaptabilidade ao PB e ao galego;

    - regras de expansão de siglas e acrónimos para PE, sua implementação e teste, com 99,88% de taxa de acerto, e discussão sobre a adaptabilidade ao PB e ao galego;
    - regras de conversão de numerais para PE, sua implementação e teste, com 99,86% de taxa de acerto, e discussão sobre a adaptabilidade ao PB e ao galego;
- Proposição de regras de desambiguação de homógrafos para 116 pares em PE e 107 pares em PB, sua implementação e teste, tendo sido obtidas as seguintes taxas de sucesso: 98,2% para PE e 97,71% para PB; discutiu-se ainda a adaptabilidade do sistema ao galego, que demonstra uma compatibilidade de pares de homógrafos com o PE na ordem dos 62,06%;
- Proposição de regras de leitura de estrangeirismos para PE, sua implementação e teste, tendo-se obtido 88,05% de taxa de acerto por palavra e 98,14% de taxa de acerto por fone, e discussão da sua adaptabilidade ao PB e ao galego;
- Proposição de regras de divisão silábica para PE, sua implementação, e teste, tendo-se obtido 99,06% de taxa de sucesso; demonstração da adaptabilidade do divisor silábico proposto para PB e para galego, tendo-se obtido 99,20% de acerto com textos em PB e 97,87% de acerto com textos em galego;
- Proposição de regras de marcação de sílaba tónica para PE, sua implementação e teste, com resultados de 99,54% de sílabas tónicas correctamente identificadas; demonstração da aplicabilidade destes algoritmos ao PB e ao galego, cujos resultados foram de 99,60% e 98,52%, respectivamente;
- Proposição de regras de transcrição grafema-fone para PE, sua implementação e teste, tendo-se obtido 99,11% de fones correctamente transcritos; desenvolvimento e implementação de dois novos algoritmos de conversão grafema-fone para PB e para galego, baseados na proposta para PE; os resultados dos testes dos dois sistemas foram de 97,44% de taxa de acerto para o PB e 98,50% para o galego.
- Integração dos algoritmos propostos com um motor de síntese baseado em HMMs e gravação de uma base de dados de fala em português.

Ao longo deste trabalho, foram sendo publicados resultados preliminares dos vários módulos aqui propostos em revistas e congressos nacionais e internacionais com revisores. Segue-se a lista das publicações relacionadas:

1. Simões, C.; Calado, A.; Braga, D.; Teixeira, C., Dias, M. 2007. "European Portuguese Accent in Non-native English models for ASR systems", *12th Iberoamerican Congress in Pattern Recognition - CIARP 2007*, Viña del Mar- Valparaíso, Chile, November 2007, pp. 738-747.

2. Braga, D.; Resende Jr.; F. G.; Marques, M. A. 2007. "Leitor de estrangeirismos para sistemas de conversão Texto-Fala em Português Europeu", *XIII Encontro Nacional da Associação Portuguesa de Linguística*, 1-3 Outubro de 2007, Évora, Portugal.

3. Braga, D.; Coelho, L.; Resende Jr., F.G.V. 2007. "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", *Proceedings of Interspeech 2007*, 27-31 Agosto de 2007, Antuérpia, Bélgica.

4. 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.

5. Braga, D.; Resende Jr., F. G. V. 2007. "Módulos de Processamento de Texto Baseados em Regras para Sistemas de Conversão Texto-Fala em Português Europeu", *Lobo, M. & Coutinho, M. A. (Orgs), XXI Encontro da Associação Portuguesa de Linguística*. Coimbra, 2-4 Outubro de 2006.pp.141-156.

6. Braga, D.; Coelho, L. 2006. "Letter-to-sound conversion for Galician TTS systems", *IV Jornadas en Tecnologías del Habla*, 8-10 de Novembro de 2006, Zaragoza, Espanha. pp. 171-176. ISBN: 84-96214-82-6.

7. Braga, D.; Freixeiro, X. 2006. "Algoritmos de Conversão Grafema-Fone em Galego para Sistemas de Conversão Texto-Fala", *VIII Congreso Internacional de Estudos Galegos,* Salvador, Bahía, Brasil, 12-15 de Setembro de 2006. (no prelo).

8. Braga, D.; Coelho, L.; Resende Jr., F. 2006. "A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese", *VI International Telecommunications Symposium (ITS2006)*, 3-6 de Setembro de 2006, Fortaleza-CE, Brasil.

9. Silva, D.; Lima, A.; Maia, R.; Braga, D.; Moraes, J. F.; Moraes, J. A.; Resende Jr. F. 2006. "A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing", *VI International Telecommunications Symposium (ITS2006)*, 3-6 de Setembro de 2006, Fortaleza-CE, Brasil.

10. Braga, D. 2006. "Grapheme-to-phone transcription algorithm for Text-to-Speech Systems in European Portuguese", *POLISSEMA – Revista de Letras do ISCAP,* nº 6, Instituto Superior de Contabilidade e Administração do Porto, Porto, Portugal.

Enquanto se pode considerar que as técnicas de síntese se encontram já num bom nível de desenvolvimento e qualidade, o mesmo não se pode dizer de certas questões ao nível do *front-end*. Além disso, enquanto que a evolução das técnicas de síntese é independente da língua, as questões relacionadas com o *front-end* assentam nas especificidades das línguas e dependem dos recursos linguísticos disponíveis. O português, apesar dos quase 20 anos de trabalho científico dos dois lados do Atlântico, apresenta ainda várias questões por resolver ou com soluções pouco

satisfatórias, designadamente ao nível da desambiguação de homógrafos, leitura de estrangeirismos e mesmo no âmbito conversão grafema-fone (por exemplo, a marcação de tonicidade).

A regularidade fonética e fonológica do português, bem como a sua ortografia de base fonológica, explicam o sucesso da aplicação da abordagem que seguimos neste trabalho, uma abordagem baseada em regras linguísticas para resolução dos problemas de processamento de texto e análise linguística na conversão texto-fala. Esta abordagem é mais económica do ponto de vista computacional, exigindo menos processamento e menos corpora de treino prévio, sendo muito adequada para tecnologias em ambiente móvel, em que a memória computacional é mais reduzida, porque o aparelho é cada vez mais pequeno e leve. A metodologia seguida e demonstrada nesta dissertação pode ser expandida a outras línguas românicas, como ficou demonstrado com o galego.

A adaptabilidade total ou implicando poucas alterações dos vários módulos apresentados ao nível do *front-end* vem a demonstrar a proximidade linguística entre o PE, PB e o galego. Por exemplo, do conjunto de regras de transcrição grafema-fone que descrevem o PE, 70,5% são totalmente aplicáveis ao PB, enquanto 21,9% apresentam igual padrão de input mas diferente padrão de output. Estes dados representam um total de 92,5% de compatibilidade entre as regras que descrevem a conversão grafema-fonema do PE e do PB. Demonstrou-se ainda que 80,2% das regras de transcrição grafema-fone que descrevem o PE se aplicam totalmente ao caso do galego e que 19,1% apresentam igual padrão de input mas diferente padrão de output. Estes dados representam um total de 95,3% de compatibilidade entre as regras que descrevem a conversão grafema-fonema do PE e do galego. Essa proximidade linguística é ainda reforçada pelos resultados dos testes dos algoritmos de divisão silábica e de marcação de sílaba tónica com textos em PE, PB e galego, que sem qualquer adaptação revelaram resultados rondando os 98 e 99%. Além disso, no domínio do inventário lexical de homógrafos, verificou-se que existe 92,24% de taxa de correspondência entre PE e PB e 62,06% entre PE e galego. Também os sub-módulos do pré-processamento são em grande parte compatíveis entre PE, PB e galego, implicando pequenas alterações apenas, exceptuando a conversão de siglas/acrónimos em galego, cujo funcionamento necessita de um estudo mais aprofundado. O único aspecto da análise fonética que parece apresentar menor compatibilidade entre PE, PB e galego é a leitura de estrangeirismos, embora a falta de estudos sobre a integração das palavras estrangeiras no PB e no galego não nos permita possuir dados definitivos em relação a este assunto. O fenómeno da integração ou não integração de estrangeirismos em PE, PB e galego requer mais investigação linguística, visto que os estrangeirismos são os principais responsáveis pelos erros nos vários módulos propostos, nomeadamente no divisor silábico e no marcador de tonicidade.

As principais limitações deste trabalho prendem-se com problemas de previsão da qualidade da vogal em radicais de formas verbais com alternância vocálica (ex. <calo, calas, cala, calam> [a] vs <calamos, calais> [6]) e em radicais de palavras derivadas por sufixação (ex. <sozinho>, <amestrado>, <honestamente>), em que o timbre vocálico é explicado por razões etimológicas apenas, impossíveis de reduzir a regras.

Perante a iminência da entrada em vigor do Acordo Ortográfico da Língua Portuguesa de 1990[219] em 2008, cumpre-nos ainda discutir as implicações que este documento pode ter nos algoritmos propostos nesta dissertação, uma vez que assentam na proposição de regras que partem da ortografia da língua. O Acordo Ortográfico de 1990 tem por objectivos defender a unidade essencial da língua Portuguesa e aumentar o seu prestígio internacional. Este acordo, celebrado em 1990, visa unificar a ortografia do português, que mantém actualmente duas normas em vigor, uma no Brasil e outra nos restantes países de língua oficial Portuguesa. Várias reacções negativas de Portugal e do Brasil têm impedido a entrada em vigor do Acordo. Mas, segundo notícia publicada na página web da cadeia de televisão portuguesa SIC online[220], o governo português prometeu ratificar o documento até ao final de 2007 para entrada em vigor em 2008. Segundo especialistas, as alterações irão afectar apenas 0,45% do vocabulário no Brasil e 1,6% das palavras escritas em Portugal. Até mesmo na Galiza houve ecos desta notícia.[221]

Em linhas muito gerais, o Acordo Ortográfico uniformiza o uso do hífen, elimina o trema no Brasil e as consoantes mudas <p> e <c> em Portugal e suprime certos acentos ortográficos (no Brasil, <enjôo>, <vôo> perdem o acento circunflexo e <heróico>, <idéia> perdem o acento agudo; em Portugal e no Brasil, as terceiras pessoas do plural do presente do indicativo de verbos como <crer>, <ver>, <ler> perdem o acento circunflexo, passando a escrever-se <creem>, <veem>, <leem>).

As alterações ao nível da supressão do hífen em palavras como <antirreligioso> ou <autoestrada> não constituem problema para os nossos módulos, uma vez que o sistema não tem limite de número de grafemas por palavra.

Outras propostas de alteração ortográfica, porém, têm impacto no actual desempenho do nosso sistema, especificamente ao nível da conversão grafema-fone.

A primeira proposta com implicação na conversão grafema-fone do PE é a supressão das consoantes <c> e <p> sempre que não sejam pronunciadas. Esta medida invalida a actuação das regras 12 do grafema <a>[222], 15 do grafema <e>[223] e 1

---

[219] Para consulta ao texto do Acordo Ortográfico da Língua Portuguesa de 1990, ver: http://www.cplp.org/docs/documentacao/Acordo%20ortogr%C3%A1fico%20retirado%20internet.pdf (01-01-2008). Sobre a polémica e várias questões relacionadas, ver Ciberdúvidas em: http://ciberduvidas.sapo.pt/pergunta.php?id=20570 (01-01-2008). Para a história da ortografia em português, veja-se ainda: http://www.portaldalinguaportuguesa.org/?action=acordo-historia (01-01-2008).

[220] Notícia de autoria da jornalista Catarina Lúcia Carvalho em 20 de Dezembro de 2007: "Ao fim de 17 anos, o Acordo Ortográfico poderá finalmente entrar em vigor em 2008. Portugal prometeu ratificar ainda este ano o documento que vai alterar a forma como escrevem milhões de falantes da língua portuguesa. Uma mudança que está longe de ser consensual." (disponível em: http://sic.sapo.pt/online/noticias/vida/20071239+-+Acordo+Ortografico+ate+2008.htm, 01-01-2008)

[221] Artigo "O idioma português aplicará unha ortigrafía máis simple e uniforme", in *La Voz de Galicia*, 13 de Dezembro de 2007.

[222] Perante o padrão gráfico: ...<a> <ct, cç, pt, cc> ..., <a> realiza-se [a] (ex. <acção, captura, faccioso>).

[223] Perante o padrão gráfico:...<e><(ct, cç, cc, gn, pç, pt)>..., <e> realiza-se [E] (ex. <dialecto, direcção>).

do grafema <p>[224]. As duas primeiras regras permitem prever a qualidade vocálica aberta dos grafemas <a> e <e> quando seguidos de <ct, cç, cc, gn, pç, pt>. A ausência das consoantes mudas torna muito difícil prever o output vocálico destas vogais, introduzindo certamente erros no actual sistema. Apenas uma lista de excepções poderá resolver de forma eficaz o vazio deixado por esta regra. Em relação à regra 1 do grafema <p>, a sua supressão resultante do Acordo não terá reflexos no desempenho do sistema, já que <p> será [p] em todos os contextos, o que está coberto pela regra *default* da conversão de <p>.

A supressão dos acentos gráficos em <pêlo> (substantivo), <pára> (verbo) e <pôr> (verbo) tornará estas palavras homógrafas de <pelo> (contracção da preposição com o artigo), <para> (preposição) e por <preposição>, o que representa mais um problema que impossibilita a previsão automática da qualidade da vogal tónica, uma vez que, na actual versão do sistema, <pelo, para, por> são palavras átonas, e por isso, transcritas com vogais fechadas. Além disso, também são identificadas como átonas pelo marcador de sílaba tónica. Adaptações do nosso sistema à nova regra trazida pelo Acordo passam pela expansão da lista de pares de homógrafos, que passará a incluir estas palavras, com novos algoritmos assentes numa desambiguação sintáctica.

A eliminação em Portugal e no Brasil do acento circunflexo das terceiras pessoas do plural do presente do indicativo de verbos como <crer>, <ver>, <ler> terá impacto na conversão grafema-fone da vogal tónica em PE, que será lida como *default* [@], mas não em PB, cujo *default* é [e]. Uma vez mais, estas palavras deverão integrar uma lista de excepções, por serem poucas mas muito frequentes.

A supressão do acento agudo em <bóia>, <heróico>, <idéia> no PB também dificulta a previsão do timbre dos grafema <o> e <e> nestes contextos. Contudo, as seguintes regras poderiam ser propostas:

- perante o padrão gráfico: …<e><ia><Pont, SP, s>…, <e> realiza-se [E] (ex. <ideia>, <geleia>);
- perante o padrão gráfico: …<o><ia, io><Pont, SP, s>…, <o> realiza-se [O] (ex. <boia>, <geleia><apoio>).
-

A supressão do acento agudo em <enjôo> e <vôo> em PB não tem qualquer efeito na conversão grafema-fone, divisão silábica ou marcação de sílaba tónica.

Em relação à ortografia brasileira, a eliminação do trema impossibilita a realização da regra 1 do <u>[225], tornando muito difícil a previsão da articulação de <u> em palavras como <lingüiça> ou <lingüista>. A solução a ser adoptada será análoga à seguida em PE, que já não possui o trema, mas que apresenta os mesmos contextos de articulação do grafema <u>. Essa solução passa pela utilização da mesma lista de excepções construída para o PE (cf. Regras 1 do grafema <q> e 2 do grafema <g>).

Em resumo, o Acordo Ortográfico implicará a reformulação de algumas regras de conversão grafema-fone do PE e do PB, mas a grande parte das regras não será afectada.

---

[224] Perante o padrão gráfico:...<p><t, ç>..., <p> realiza-se [ ] (ex. <óptimo>).

[225] Perante o padrão gráfico:...<...<ü>... , <u> realiza-se [w] (ex. <lingüística>).

No que respeita às aplicações possíveis deste trabalho, além da natural integração num sintetizador de fala, com todas as aplicações já referidas, desde a acessibilidade à mobilidade, gostaríamos de destacar as seguintes:

- conversão fonética, divisão silábica e marcação de tonicidade automáticas de léxicos e dicionários electrónicos, que são recursos muito dispendiosos e escassos em português, e que constituem a base de muitas ferramentas linguísticas, como sintetizadores de fala e reconhecedores de voz que usam outras abordagens para a análise linguística;
- conversão fonética automática de textos que sirvam de corpora de treino para o aperfeiçoamento de outras técnicas de processamento da linguagem natural;
- desenvolvimento de software educativo para o ensino da estrutura fonética, fonológica e sintáctica do português;
- desenvolvimento de software clínico, destinado à terapia da fala, através de exercícios que trabalhem a relação grafema-fone e a estrutura fonológica da língua.

Como trabalho futuro pretendemos avaliar a compatibilidade entre o nosso algoritmo e outros algoritmos de conversão texto-fala para outras variedades do português. É ainda nossa intenção submeter os vários módulos do nosso sistema a avaliação, permitindo assim compará-los com outras técnicas, tendo em consideração os mesmos parâmetros de análise. A expansão dos módulos de desambiguação de homógrafos está em curso, bem com a sua extensão ao galego. Pretendemos ainda expandir o leitor de estrangeirismos, o conversor de siglas e acrónimos e o conversor de numerais para PB e galego Está também em curso o desenvolvimento de uma página web em que o utilizador poderá usar e testar os algoritmos aqui propostos com textos e frases à escolha.

Referências bibliográficas

Albano, E., Moreira, A. 1996. "Archisegment-based Letter-to-Phone Conversion for Concatenative Speech Synthesis in Portuguese", *Proceedings of ICSLP 96 – Fourth International Conference on Spoken Language*. Philadelphia, USA. Volume 3. pp. 1708 – 1711.

Alcalá, X. 2006. "Substituír e non inventar", *AA.VV. Lingua e Investigación. II Xornadas sobre Lingua e Usos*. A Coruña: Universidade da Coruña.

Allen, J., Hunnicut, M. S., and Klatt, D. H. 1987. *From Text to Speech: The MITalk system*. Cambridge: Cambridge University Press.

Allen, J., Hunnicutt, S. and Klatt, D. 1987. *From text to speech : the MITalk system*. Cambridge: Cambridge University Press.

Alvarez Blanco, R. 1991."O sistema fonolóxico do galego. Comparación co do portugués", *D. Kremer (ed.), Actes du XVIIIème Congrès International de Linguistique et Philologie Romanes*. Tübingen: Max Niemeyer Verlag. pp. 517-530.

Andrade, A. R., Lopes, A. L. 2003. "O tratamento dos estrangeirismos nas últimas edições do Dicionário da Língua Portuguesa da Porto Editora", *Revista de Lexicografia da Universidada da Coruña*. vol. IX. A Coruña: Universidade da Coruña, pp. 7-28.

Baggia, P.; Badino, L.; Bonardo, D.; Massimino, P. 2006. "Achieving perfect TTS intelligibility". Originally presented at the AVIOS Technology Symposium, SpeechTEK West 2006. Loquendo White Paper. Available at: http://www.loquendo.com/en/company/whitepapers.htm.

Baggia, P.; Mosso, S. 2005. Speech Technologies and Multimodality: the solution for new advanced services. Loquendo White Paper. Available at: http://www.loquendo.com/en/company/whitepapers.htm.

Bailly, G.; Campbell, N.; Mobius, B. 2003. "ISCA Special Session: Hot topics in Speech Synthesis",*Eurospeech 2003*, Geneva, Suíça. pp. 37-40.

Barbosa, F. *et al.* 2003. "Grapheme-Phone Transcription Algorithm for a Brazilian Portuguese TTS", *Mamede et al. (eds): Proceedings of PROPOR 2003*. Heidelberg: Springer Berlin. pp. 23-30.

Barbosa, F.L.; Rosa, M.C.; Gonçalves, C.A.; Resende Jr., F.G. V.; 2003b. "Algoritmo para leitura de siglas em um sintetizador", *Anais do XX Simpósio Brasileiro de Telecomunicações.* Rio de Janeiro: IME/PUC-Rio. pp. 672-675.

Barbosa, F.; Ferrari, L.; Resende Jr., F. 2003c. "A methodology to analize homographs for a Brazilian Portuguese TTS system", *Mamede et al. (eds): Proceedings of PROPOR 2003*. Heidelberg: Springer Berlin. Heidelberg: Springer-Verlag. pp.57-61.

Barbosa, F.; Ferrari, L.; Resende Jr., F. G. 2003d."A distinção entre homógrafos heterófonos em sistemas de conversão texto-fala", *Silva et al. (Org.). Linguagem, Cultura e Cognição: Estudos de Linguística Cognitiva.* Coimbra: Almedina.

Barbosa, P. A. 2006. *Incursões em torno do ritmo da fala*. Campinas: Pontes.

Barbosa, P.; Violaro, F.; Albano, E.; Simões, F.; Aquino, P.; Madureira, S.; Françozo, E. 1999. "Aiuruetê: A High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythim Production", *Eurospeech'99 - 6th European Conference on Speech Communication and Technology*. Budapest, Hungria. Volume 5, pp. 2059-2062.

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

Barros, M. J.; Maia, R.; Tokuda, K.; Resende, F.; Freitas, D. 2005. "HMM-based European Portuguese TTS System", *Proceedings of Interspeech 2005*. Lisboa, Portugal. pp. 2581-2584.

Barros, M. J. 2001. Estudo Comparativo e Técnicas de Geração de Sinal para a Síntese da Fala. Dissertação de Mestrado. Universidade do Porto.

Bennett, C.L. and Black, A.W. 2006. "The Blizzard Challenge 2006", *Blizzard Challenge Workshop, satellite event of Interspeech 2006 – ICSLP*. Pittsburgh, USA.

Bergström, M & Reis, N. 2007. *Prontuário ortográfico e guia da língua portuguesa*. Cruz Quebrada: Casa das Letras.

Beskow, J. 2003. *Talking heads: Models and applications for multimodal speech synthesis*. PhD Thesis, Centre for Speech Technology – KTH. Stockholm,Sweden.

Bick, E. 2000. The parsing system "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD Thesis. Aarhus: Aarhus University Press.

Black, A. and Tokuda, K. 2005. "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases", *Proceedings of Interspeech 2005*. Lisbon, Portugal. pp. 77-80.

Black, A. and Lenzo, K. 2004. "Multilingual Text-to-Speech Synthesis", *Proceedings of ICASSP 2004*, Montreal, Canada. volume 1, pp. 373–376.

Black, A.W., Lenzo, K. and Pagel, V. 1998. "Issues in building general letter to sound rules", *Third ESCA/COCOSDA Workshop on Speech Synthesis.* Jenolan Caves House, Blue Mountains, Australia. pp.77-80

Boeffard, O.; Violaro, F. 1994. "Using a Hybrid Model in a Text-To-Sppech System to Enlarge Prosodic Modifications", *Proceedings of ICSLP 94*. Yokohama, Japan. pp.727-730.

Bonafonte, A.; Hoge, H.; Tropf, H; Moreno, A.; Heuvel, H.; Sündermann, D.; Ziegenhain, U.; Pérez, J.; Kiss, I. 2004. *TTS Baselines and Specifications.* Deliverable no.D8 do Projecto TC-Star disponível em: http://www.tc-star.org/.

Bosch, A. & Daelemans, W. 1993. "Data-oriented methods for grapheme-to-Phoneme conversion", *Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL'93).* Utrecht, Holanda. pp. 45-53.

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.

Braga, D. 2006a. "Grapheme-to-phone transcription algorithm for Text-to-Speech Systems in European Portuguese", *POLISSEMA – Revista de Letras do ISCAP*, vol. nº 6, Instituto Superior de Contabilidade e Admnistração do Porto, Porto, Portugal. pp. 109-124.

Braga, D., Coelho, L.; Resende Jr., F. G. V. 2006b. "A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese", *VI International Telecommunications Symposium (ITS2006)*. Fortaleza-CE, Brasil. pp. 328-333.

Braga, D.; Coelho, L. 2006c. "Letter-to-sound conversion for Galician TTS systems", *IV Jornadas en Tecnologías del Habla*. Zaragoza, Espanha. pp. 171-176.

Braga, D.; Coelho, L.; Resende Jr., F. G. V. 2007a. "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", *Proceedings of Interspeech 2007*. Antuérpia, Bélgica. pp. 1761-1764

Braga, D.; Coelho, L.; Resende Jr., F. G. V., Dias, M. S. 2007b. "Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality", *Advances in Speech Technology, 14th International Workshop*. Maribor, Eslovénia.

Braga, D.; Freixeiro, X. 2006. "Algoritmos de Conversão Grafema-Fone em Galego para Sistemas de Conversão Texto-Fala", *VIII Congreso Internacional de Estudos Galegos*, Salvador, Bahía, Brasil. (no prelo).

Braga, D.; Marques, M. A. 2004. "The Pragmatics of the prosodic features in the political debate", *Proceedings of the International Conference of Speech Prosody 2004*, 23-26 Março 2004, Nara, Japão. pp. 321-324.

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

Braga, D.; Resende Jr.; F. G.; Marques, M. A. 2007d. “Leitor de estrangeirismos para sistemas de conversão Texto-Fala em Português Europeu”, *XIII Encontro Nacional da Associação Portuguesa de Linguística*, Évora, Portugal.

Braga, D.; Coelho, L.; Resende Jr., F.G.V.; Dias, M. S. 2007e. “Subjective and Objective Assessment of TTS Voice Font Quality”, *XII International Conference Speech and Computer - SPECOM 2007*. Moscovo, Rússia. pp. 306-311.

Bulut, M.; Narayanan, S. S., and Syrdal, A. K. 2002. “Expressive speech synthesis using a concatenative synthesizer”, *Proceedings of ICSLP 2002*, Denver, USA. pp. 79-84.

Cabral, J. 2006. *Emotive Speech Synthesis*. Ms. Thesis. IST/INESC-ID, Lisboa.

Cabral, J.; Oliveira, L.C. 2006. “EmoVoice: a System to Generate Emotions in Speech”, *Proceedings of Interspeech 2006*. Pittsburgh, USA. pp. 1798-1801.

Cahn, J. E. 1990. “The generation of affect in synthesized speech”, *Journal of the American Voice I/O Society*, Vol. 8, pp. 1–19.

Callou, D. & Leite, I. 2002. *Como falam os brasileiros*. Rio de Janeiro: Jorge Zahar Editor.

Campillo Díaz, F.; Rodríguez Banga, E. 2005. "Evaluación del modelado acústico y prosódico del sistema de conversión texto-voz Cotovía", *Procesamiento del Lenguaje Natural*, núm. 35, pp. 5-12.

Campillo, F.; Santen, J. Banga, E. 2006. “A model for the f0 reset in corpus-based intonation approaches”, *Proceedings of Interspeech 2006.* Pittsburgh, USA, pp. 2362-2365.

Carballeira Anllo, X.M. (coord.) 2000. *Gran diccionario Xerais da lingua galega*, Vigo: Xerais.

Carvalho, P.; Trancoso, I.; Oliveira, L. C. 2003. “WFST based Unit Selection for Concatenative Speech Synthesis in European Portuguese”, *Proceedings of ICPhS'2003 - 15th International Congress of Phonetic Sciences*. Barcelona, Spain. Pp 2333-2336.

Caseiro, D. A., Trancoso, I. “Grapheme-to-Phone Using Finite-State Transducers”, *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. Santa Monica, USA.

Caseiro, D. A.; Trancoso, I.; Viana, M. Céu; Barros, M. 2003. “A Comparative Description of GtoP modules for Portuguese and Mirandese using Finite State Transducers”, *Proceedings of ICPhS'2003 - 15th International Congress of Phonetic Sciences*, Barcelona, Spain. Pp 2605-2608.

Casteleiro, J. M. (coord.) 2001. Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa. 2 vols. Lisboa: Editorial Verbo.

Castro, Ivo. 1991. *Curso de História da Língua Portuguesa*. Lisboa: Universidade Aberta.

Céu Viana, M.; Trancoso, I.; Silva, F. H. 1994. “On the pronunciation of proper names and acronyms in European Portuguese”, *2nd Onomastica Research Colloquium*, London, United Kingdom.

Chen, Y.; You, J.; Chu, M.; Zhao, Y. and Wang, J. 2006. "Identifying language origin of person names with N-grams of different units", *Proceedings of ICASSP2006*. Toulouse, France. pp 729-732.

Chotimongkol, A. & Black, A. 2000. “Statistically trained orthographic to sound models for Thai”, *Proceedings of ICSLP2000*. Beijing, China. Volume 2, 551-554

Cintra, L. (1971) 1995. “Nova Proposta de Classificação dos dialectos galego-portugueses”, *Estudos de Dialectologia Portuguesa*. Lisboa: Sá da Costa.

Cirigliano, R. J. R., Monteiro, C., Barbosa, F. L., Resende Junior, F. G. V., Couto, L. R., Moraes, J. A., 2005. “Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos”, *Anais do XXII Simpósio Brasileiro de Telecomunicações (SBrT 2005).* Campinas, Brazil, pp. 544-549.

Coelho, L. 2005. *Etiquetagem Automática de Sinais de Fala – Anotação e Segmentação*, M.S. thesis, Faculdade de Engenharia da Universidade do Porto.

Coelho, L.; Braga, D., Barros, M., Freitas, D. 2004. “Na ponta da Língua: Uma nova forma de acesso à informação”, *Actas da Conferência da Associação Portuguesa de Sistemas de Informação 2004*, Lisboa, Portugal.

Coker, Cecil H.; Church, Kenneth W.; Liberman, Mark Y. 1990. “Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis”, *Proceedings of the ESCA Workshop on Speech Synthesis*. Autrans, France. pp. 83-86.

Coker, Cecil H.; Kenneth, W. Church and Liberman, Mark Y. “Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis”, *Conference on Speech Synthesis. European Speech Communication Association*, 1990.

Costa, F. A da. 1990. *Dicionário de Estrangeirismos*. Lisboa: Editorial Domingos Barreira.

Cuesta, P.; Luz, M. 1971.*Gramática da Língua Portuguesa*. Lisboa: Edições 70, pp. 281-331.

Cunha, C. & Cintra, L. 1992. *Nova gramática do português contemporâneo*. Lisboa: Sá da Costa.

Damper, R.I.; Marchand Y.; Adamson, M.J.; Gustafson, K. 1998 “Comparative Evaluation of Letter-To-Sound Conversion Techniques For English Text-To-Speech Synthesis”, *Proceedings of 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, Jenolan Caves, Australia, pp. 53-58.

De Martino, J. M.; Magalhães, L. P.; Violaro, F. 2006. "Facial Animation Based on Context-Dependent Visemes", *Computers & Graphics*, Amsterdam: Elsevier. Volume 30, Issue 6. pp. 971-980.

Deller Jr., J.; Hansen, J.; Proakis, J. 2000. *Discrete-Time Processing of Speech Signals.* Piscataway, NJ: IEEE Press.

Dutoit, Thierry. 2001. *An Introduction to Text to Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Eide, E.M., Bakis, R., Hamza, W., and Pitrelli, J. F. 2004. "Towards synthesizing expressive speech", *Narayanan, S. S. and Alwan, A. (Eds.), Text to Speech Synthesis: New paradigms and Advances*. New Jersey: Prentice Hall.

Erro, D.; Moreno, A. 2007. "Frame Alignment Method for Cross-Lingual Voice Conversion", *Proceedings of Interspeech 2007*. Antwerpen, Belgium. pp. 1969-1972.

Estrela, E.; Soares, M. A.; Leitão, M. J. 2004. *Saber Escrever, Saber Falar*. Lisboa: Dom Quixote.

Estrela, E.; Soares, M. A.; Leitão, M. J. 2004. Saber escrever. Saber falar. Um guia completo para usar correctamente a língua portuguesa. Lisboa: Dom Quixote.

Fernández Rei, F. 1990. *Dialectoloxía da Lingua Galega*. Vigo: Xerais.

Ferrari, L ; Barbosa, F. ; Resende Jr., F. G. V. 2003. "Construções gramaticais e sistemas de conversão texto-fala: o caso dos homógrafos". *Proceedings of the International Conference on Cognitive Linguistics*. Braga.

Ferreira, A. B. H.; Ferreira, M. B. Anjos, M. (coord e ed). 2004. *Novo dicionário Aurélio da língua portuguesa*. 3.ed. Curitiba: Positivo.

Ferreira, H. e Freitas, D. 2004. "Enhancing the Accessibility of Mathematics for Blind People: The AudioMath Project", *Proceedings of ICCHP2004*. Paris, France. LNCS 3118, Springer. pp. 678-685.

Ferreira, H. e Freitas, D. 2005. "AudioMath: Towards Automatic Readings of Mathematical Expressions", *Human-Computer Interaction International (HCII) 2005.* Las Vegas, USA.

Ferreira, H. 2005. *Leitura Automática de Expressões Matemáticas - AudioMath*. Dissertação de Mestrado. Faculdade de Engenharia da Universidade do Porto.

Ferreira, M. Carrilho, E.; Lobo, M., Saramago, J.; Segura, L. 1995. "Variação linguística: perspectiva dialectológica", *Faria, et al. Introdução à Linguística Geral e Portuguesa.* Lisboa: Caminho.

Ferreira, M.; Carrilho, E.; Saramago, J.; Cruz, L. 1996. "Variação Linguística: perspectiva dialectológica", *Faria, I. et al. 1996. Introdução à Linguística Geral e Portuguesa*. Lisboa: Caminho.

Fraser, M.; King, S. 2007. "The Blizzard Challenge 2007", *Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany.

Freitas, T.; Ramilo, M. C. e Soalheiro, E. 2003. "O processo de integração dos estrangeirismos no Português Europeu", *Mendes & Freitas (orgs.) Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa, Portugal.

Freixeiro Mato, X. R. 2006. *Gramática da Lingua Galega I. Fonética e Fonoloxía*. Vigo: Edicións a Nosa Terra.

Freixeiro Mato, X. R. 2006a. *Gramática da Lingua Galega III – Semántica*. Vigo: Edicións a Nosa Terra.

Freixeiro Mato, X. R. 2006b. *Manual de Gramática Galega*. Vigo: Edicións a Nosa Terra.

García González, C. & González González, M. (orgs.) 1997. *Diccionario da Real Academia Galega*. A Coruña: Real Academia Galega.

Giangola, James P. 2001. *The pronunciation of Brazilian Portuguese*. Muenchen: Lincon Europa.

Gomes, L.C.T. 1998. Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras. Dissertação de Mestrado. Campinas: Unicamp.

González González, M. 2004."A síntese de voz en lingua galega: o proxecto Cotovía", *Revista Galega do Ensino*, núm. 44, pp. 199-215.

Grande Dicionário da Língua Portuguesa da Porto Editora. 2004. 1ª edição. Porto: Porto Editora.

Hamza, W.; Bakis, R.; Eide, E.M.; Picheny, M. A.; and Pitrelli, J. F. 2004. "The IBM expressive speech synthesis system", *Proceedings of ICSLP 2004*, Jeju, Korea. pp. 1099- 1108.

Zen, H.; Nankaku, Y.; Tokuda, K.; Kitamura, T. 2006. ``Speaker adaptation of trajectory HMMs using feature-space MLLR," *Proceedings of Interspeech 2006 – ICLSP*. Pittsburgh, USA. pp.1141-1144.

Houaiss, A. (coord.) 2001. *Dicionário Houaiss da Língua Portuguesa*. Rio de Janeiro: Objectiva.

Huang, C. B., Son-Belt, M. A. and Baggett, D. M. "Generation of pronunciation from orthographies using transformation-based error-driven learning", *Proceedings of ICSLP94*. Yokohama, Japan. pp. 411-414.

Huang, X.; Acero, A. and Hon, H.W. 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. New Jersey: Prentice Hall.

Iriarte Sanromán, Á. 2001. *A Unidade Lexicográfica. Palavras, colocações, frasemas, pragmatemas*. Braga: Centro de Estudos Humanísticos da Universidade do Minho.

Jurafsky, D.; Martin, J. 2007. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. (draft version; last updated in October 2007). Disponível em: http://www.cs.colorado.edu/~martin/slp2.html.

Kaplan, R. M. & Kay, M. 1994. "Regular models of phonological rule systems", *Computational Linguistics 20(3)*, pp. 331-378.

Keller, E. 2002. *Improvements in Speech Synthesis: Cost 258: The Naturalness of Synthetic Speech*. New York: Wiley.

Klatt, D.. 1987. "Review of text-to-speech conversion for English", in *The Journal of the Acoustical Society of America*, Volume 82, Issue 3. pp.737-793.

Laporte, E. 1993. *Phonetique et transducteurs. Technical report*, Paris: Université Paris 7.

Lavouras Lopes, A. e Rebello d'Andrade, A. 1997. "Primeira fase da instalação do estrangeirismo", *Actas do XIII Encontro da APL*. Colibri: Lisboa.

Lavouras Lopes, A. 1992. *Os estrangeirismos no português contemporâneo*. Texto Policopiado.

Lee, S.; Bresch, E.; Adams, J.; Kazemzadeh, A.; and Narayanan, S. S. 2006. "A study of emotional speech articulation using a fast magnetic resonance imaging technique", *Proceedings of ICSLP 2006*, Pittsburgh, USA.

Lemmety, S. 1999. *Review of Speech Synthesis Technology*. Ms Thesis. Helsinki University of Technology.

Llisterri, J.; Martí Antonín, M. A. 2002. *Tratamiento del Lenguaje Natural*. Barcelona: Edicions de la Universitat de Barcelona, S.L. Unipersonal.

Llitjos, A.F. and Black, A.W. 2001. "Knowledge of language origin improves pronunciation accuracy of proper names", *Proceedings of Eurospeech 2001*, Alborg, Denmark.

Real Academia Galega/ Instituto da Língua Galega. 2003. *Normas ortográficas e morfolóxicas do idioma galego*. Vigo: Real Academia Galega/ Instituto da Língua Galega.

Losada Soto, R. M. 2004. "Unha adaptación do SAMPA para a Lingua Galega", *A Lingua Galega, Historia e Actualidade – Actas do I Congresso Internacional*. Conselho da Cultura Galega, Instituto da Lingua Galega. pp. 615-625.

Lucassen, J. M. & Mercer, R. L. 1984. "Discovering Phonemic based forms automatically: an information theoretic approach", *IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 42.5.1-42.5.4.

Lucassen, J. M. 1983. *Discovering phonemic base forms automatically: an information theoretic approach*. Technical Report RC 9833 (#43527), IBM T.J., Watson Research Center.

Lyons, J. 1977. *Semantics*. 2 vols. Cambridge: Cambridge University Press.

Machado, J. P. 1994. *Estrangeirismos na Língua Portuguesa*. Lisboa: Editorial Notícias.

Maia, R.; Toda, T.; Zen, H.; Nankaku, Y.; Tokuda, K. 2007. “A Trainable Excitation Model for HMM-Based Speech Synthesis”, *Proceedings of Interspeech 2007*. Antwerpen, Belgium, pp. 1909-1912.

Maia, R.; Zen, H.; Tokuda, K.; Kitamura T. and Resende, Jr., F. G. V. 2006. “An HMM-based Brazilian Portuguese speech synthesizer and its characteristics", *IEEE Journal of Communication and Information Systems*. No. 2, vol. 21, pp. 58-71.

Maia, R.; Zen, H.; Tokuda, K.; Kitamura, T. and Resende, F. G. 2003. “Towards the development of a Brazilian Portuguese text-to-speech system based on HMM”, *Proceedings of the European Conferece on Speech Communication and Technology EUROSPEECH* 2003. Geneva, Switzerland. pp. 2465–2468.

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

Mao, X.; Dong, Y.; Han, J.; Huang, D.; Wang, H. 2007. “Inequality Maximum Entropy Classifier with Character Features for Polyphone Disambiguation in Mandarin TTS Systems”, *Proceedings of ICASSP 2007*, Honolulu, Hawai. pp. IV-705-IV-708.

Mareuil, B. *et al.* 2005. "Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters", *Proceedings of Interspeech 2005*. Lisbon, Portugal.

Mariño, J. B., Nadeu, C., Llisterri, J. 1987. “Síntesis automática del habla”, *Mompín (coord..) Inteligencia Artificial: conceptos, técnicas y aplicaciones*. Barcelona: Marcombo (Serie mundo electrónico, 13). pp. 157-165.

Martínez, M. 1997. El Paradigma Emergente. Hacia una nueva teoría de la racionalidad humana.2 ed. Editorial Trillas, México.

Martínez, M. 2006. “Nuevo Paradigma Epistemológico de la Ciencia”, *ConcienciActiva 21*, número 14, Octubre 2006. Caracas: Fundación ConcienciActiva. pp. 15-59.

Martino, J. M.. 2005. Animação Facial Sincronizada com a Fala: Visemas Dependentes do Contexto Fonético para o Português do Brasil. Tese de Doutoramento. DCA/FEEC/UNICAMP, 2005.

Martins, M. R. D. 1998. *Ouvir Falar. Introdução à Fonética do Português*. Lisboa: Caminho.

Mateus, M. & Andrade, E. 2000. *The Phonology of Portuguese*. Oxford: Oxford University Press.

Mateus, M. H. M. 1975. *Aspectos da Fonologia Portuguesa*. Lisboa: Centro de Estudos Filológicos.

Mateus, M. H. M., Andrade, A., Viana, M. C., Villalva, A. 1990. *Fonética, Fonologia e Morfologia do Português*. Lisboa: Universidade Aberta.

Mendes, H.M.; Oliveira, C.; Teixeira, A. 2004. "PLE: uma sigla para ler ou soletrar?", *Cadernos de PLE/ Centro de Línguas e Culturas*, nº 3, 2003 (2004), Universidade de Aveiro. pp. 121-139.

Meng, H. M.; Seneff, S. ; Zue, V. 1994. "Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation", *ARPA Human Language Technology Workshop*. Princeton, USA.

Morais, E. S. 2006. Algoritmos OPWI e LDM-GA para Sistemas de Conversão Texto-Fala de Alta Qualidade Empregando a Tecnologia SCAUS. Dissertação de doutoramento. DECOM/FEEC/UNICAMP.

Nakamura, K.; Toda, T.; Saruwatari, H.; Shikano, K. 2007. "Impact of Various Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees", *Proceedings of Interspeech 2007*. Antwerpen, Belgium. pp. 2517-2520.

Nicodem , M. V.; Seara, R.; Pacheco, F. S. 2005. "Reducing the Natural Click Effect within Database for High Quality Corpus-Based Speech Synthesis", *8th IEEE International Symposium on Signal Processing and its Applications*. Sydney, Austrália. pp. 607-610.

Nicodem; M. V. , Kafka , S. G.; Seara Junior , R.; Seara, R. 2007. "Refinamento da Segmentação Fonética em Aplicações de Síntese de Fala", *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007*). pp.1-6.

Nicola, J.; Terra, E.; Menón, L. 2003. *1001 estrangeirismos de uso corrente em nosso cotidiano*. São Paulo: Editora Saraiva.

Nogueira, R. Sá. 1994. *Dicionário de Verbos Portugueses Conjugados*. Lisboa: Clássica Editora.

Oliveira, C.; Moutinho, L.; Teixeira, A. 2004. "Um novo sistema de conversão grafema-fone para PE baseado em transdutores", *Actas do II Congresso Internacional de Fonética e Fonologia*, Maranhão, Brasil (no prelo).

Oliveira, C.; Moutinho, L.; Teixeira, A. 2005. "On European Portuguese Automatic Syllabification", *Proceedings of Interspeech 2005*. Lisboa. Portugal.

Oliveira, C.; Moutinho, L.; Teixeira, A. 2007. "On European Portuguese Automatic Syllabification", *González González, et al.* (coords), *III Congreso Internacional de Fonética Experimental*, Santiago de Compostela: Xunta de Galicia. pp. 461-473.

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

Oliveira, L. C.; Viana, M. C.; Trancoso; I. M. 1992. "A Rule-Based Text-to-Speech System for Portuguese", *Proceedings of ICASSP'92 - International Conference on Acoustics, Speech, and Signal Processing*. San Francisco, USA.

Oliveira, L.; Viana, M. C.; Mata, A. I. and Trancoso, I. 2001. Progress report of project dixi+: A portuguese text-to-speech synthesizer for alternative and augmentative communication. Technical report, FCT.

Oliveira, L. C. 1996. *Síntese de Fala a Partir de Texto*. Dissertação de Doutoramento. Universidade Técnica de Lisboa.

Paiva, S.; Moutinho, L.; Teixeira, A. 2005. "Síntese por concatenação em variantes regionais – o falar do Porto", *Duarte, I.; Leiria, I. (org.), Actas do XX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri. pp. 777-788.

Paulo, S. G.; Oliveira, L. C. 2005. "Generation of Word Alternative Pronunciations Using Weighted Finite State Transducers", *Proceedings of Interspeech 2005*. Lisboa, Portugal.

Percybrooks, W. Moore Il, E. 2007. "New Algorithm for LPC Residual Estimation from LSF Vectors for a Voice Conversion System", *Interspeech 2007*, Antwerpen, Belgium, pp. 1977-1980.

Pollet, V.; Coorman, G. 2004. "Statistical Corpus-Based speech Segmentation", *Proceedings of Interspeech 2004*. Jeju, Korea. pp. 1929-1932.

Rabiner, L.R.; Schafer, R. W. 1978. *Digital Processing of Speech Signals.* New Jersey: Prentice Hall.

Raimundo, G.; Cabral, J. Melo, C.; Oliveira, L.; Paiva, A.; Trancoso, I. 2007. "Telling Stories with a Synthetic Character: Understanding Inter-modalities Relations", *COST Action 2102 International Workshop on Verbal and Nonverbal Communication Behaviours*. Heidelberg: Springer. pp. 310-323.

Ramos, E. (org.). s/d. *Os Lusíadas de Luís de Camões*. Porto: Porto Editora.

Rebello d'Andrade, A. e Lavouras Lopes, A. 2003. "O tratamento dos estrangeirismos nas duas últimas edições do Dicionário da Língua Portuguesa, da Porto Editora" *Revista de Lexicografía*, vol. IX. A Coruña: Universidade da Coruña, pp. 7-28.

Regueira, X. L. 1997. "Elementos para a definición dun modelo fonético estándar da lingua galega", *B. Fernández Salgado, Actas do IV Congreso Internacional de Estudios Galegos*. Vol. 1. Oxford: Centro de Estudios Galegos. pp. 179-194.

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

Ribeiro, R.; Oliveira, L. C.; Trancoso, I. 2002. "Morphossyntactic Disambiguation for TTS Systems", *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Spain. Volume V. pp. 1427-1431.

Roche E. & Schabes, Y. 1995. *Exact Generalization of Finite-State Transductions: Application to Grapheme-to-Phoneme Transcription*. Technical Report TR-95-08, Mitsubishi Electric Research Laboratories. Cambridge, USA.

Rodrigues, M. C. 2003. *Lisboa e Braga: Fonologia e Variação*. Lisboa: FCT/MCT.

Rodríguez Banga, E.; García Mateo, C.; Fernández Salgado, X. 2001. "Concatenative Text-to-Speech Synthesis based on Sinusoidal Modelling" *Keller (org.), Improvements in Speech Synthesis*. John Wiley and Sons, Ltd. pp. 39-51.

Rodríguez Río, X. A. 2003. *Metodoloxía do traballo terminográfico puntual en lingua galega*. Santiago de Compostela: Consello da Cultura Galega, Sección de Lingua.

Rodríguez Río, X. A. 2004. "O tratamento dos empréstimos: unha proposta de actuación", *A Lingua Galega, Historia e Actualidade – Actas do I Congresso Internacional*. Conselho da Cultura Galega, Instituto da Lingua Galega. pp. 407-415.

Santos, B. S. 2001. *Um discurso sobre as ciências*. Porto: Edições Afrontamento. 12ª Edição.

Santos, D. 2001. "Processamento da linguagem natural: uma apresentação através das aplicações", *Ranchhod (org.) 2001. Tratamento das Línguas por Computador. Uma introdução à Linguística Computacional e suas aplicações.* Lisboa: Caminho.

Schmidt-Radefelt, J. e Schurig, D. 1997. *Dicionário dos Anglicismos e Germanismos na Língua Portuguesa*. Frankfurt am Main: Verlag Teo Ferrer de Mesquita.

Schroder, M. 2006. "Expressing degree of activation in synthetic speech". *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 1128–1136.

Seara Jr., R.; Kafka, S.; Seara, I.; Pacheco, F.; Klein, S.; Seara, R. 2004. "Parâmetros Lingüísticos Utilizados para a Geração Automática de Prosódia em Sistemas de Síntese de Fala", *XXI Simpósio Brasileiro de Telecomunicações - SBrT 2004*. Belém, PA, Brasil. pp. 1-6,

Seara, I. C.; Nicodem, M. V.; Seara, R.; Seara Junior, R. 2007. "Classificação Sintagmática Focalizando a Síntese de Fala: Regras para o Português Brasileiro", *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007).* pp. 1-6.

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", *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", *Revista da Sociedade Brasileira de Telecomunicações*. vol. 17, nº 1, pp. 79-85.

Segura, L.; Saramago, J. 2001. "Variedades dialectais portuguesas", Mateus (org.), *Caminhos do Português*. Lisboa: Biblioteca Nacional.

Sejnowski, T. J. and Rosenberg, C. R. 1987. "Parallel networks that learn to pronounce English Text", *Complex Systems, 1*, pp. 145-168.

Shalonova, K; Tucker, R. 2003. "South Asian Languages in Multilingual TTS-Related Database", Technical Report.

Shopp, L. 2007. “TTS is finding its way”, *Speech Technology*, November/December 2007, volume 12, nº 9.

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

Simões, C.; Calado, A.; Braga, D.; Teixeira, C., Dias, M. 2007. “European Portuguese Accent in Non-native English models for ASR systems”, *12th Iberoamerican Congress in Pattern Recognition - CIARP 2007*, Viña del Mar- Valparaíso, Chile. pp. 738-747.

Simões, F. O., Violaro, F., Barbosa, P. A. e Albano, E. C. 2000. “Um Sistema de Conversão Texto-Fala para o Português Falado no Brasil”, *Revista da Sociedade Brasileira de Telecomunicações*. Vol. 15, no 2, pp. 70-77, dezembro/2000.

Soto Andión, X. & Vidal Miexón, A. 1997. “Consideracións arredor do Galego no interior da província de Pontevedra”, *Actas do I Simposio Internacional sobre o Bilingüismo*, Vigo.

Sundermann, D. Hoge, H. Bonafonte, A. Ney, H. Black, A. Narayanan, S. 2006. “Text-Independent Voice Conversion Based on Unit Selection”, *Proceedings of ICASSP 2006*. Toulouse, France. Volume 1, pp. I-I.

Tarrío Barreiro, A.; Seoane García, M. 1997. “Contacto e interferências lingüística na prensa escrita”, *Actas do I Simposio Internacional sobre o Bilingüismo*, Vigo.

Taylor, P. 2005. “Hidden Markov Models for Grapheme to Phoneme Conversion”, *Proceedings of Interspeech 2005*, Lisbon, Portugal. pp .1973-1976.

Taylor, Paul. 2007. *Text-to-Speech Synthesis*. Draft version available at: http://mi.eng.cam.ac.uk/~pat40/book.html. To be published in 2008 by Cambridge University Press.

Teixeira, A., Oliveira, C., Moutinho, L. 2006a. “Machine Learning of European Portuguese Grapheme-To-Phone Conversion using a Richer Feature Set”, *Revista do DETUA*, Vol. I, nº 1, Aveiro.

Teixeira, A., Oliveira, C., Moutinho, L., 2006b. “On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme- Phone Conversion”, *R. Vieira, P. Quaresma, Maria G. V. Nunes, N. Mamede, C. Oliveira, M. C. Dias (Eds), Computacional Processing of the Portuguese Language (Proceedings 7th InternationalWorkshop, PROPOR 2006)*, Springer. pp. 212-215.

Teixeira, A. 2000. *Síntese Articulatória das vogais nasais do Português Europeu*. Dissertação de Doutoramento. Universidade de Aveiro.

Teixeira, J.P.; Freitas, D.; Braga, D.; J. P.; Barros; M. J.; Latsh, V. “Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB”, *Proceedings of Eurospeech 2001 – Scandinavia*, Denmark, September 2001. pp. 1707-1710.

Teixeira, J. P. 2004. *A Prosody Model to TTS Systems*. PhD Thesis, Faculdade de Engenharia da Universidade do Porto.

Teixeira, J.P. 1995. *Modelização Paramétrica de Sinais Para Aplicação em Sistemas de Conversão Texto-Fala*. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto.

Teixeira, J. P.; Freitas, D. 1998. "MULTIVOX- Conversor Texto-Fala para Português", *Proceedings of PROPOR'98*, Porto Alegre, Brasil.

Teixeira, J. P.; Gouveia, P., Freitas, D. 2000. "Divisão silábica automática do texto escrito e falado", *Proceedings of PROPOR'2000*. Atibaia, SP. Brasil.

Terminologia Linguística para os Ensinos Básico e Secundário – TLEBS, Diário da República, 24 de Dezembro. Portaria 1488/2004. Disponível em: http://www.dgidc.min-edu.pt/TLEBS/Portaria1488%2024Dez2004.pdf

Tesprasit, V.; Charoenpornsawat, P. and Sornlertlamvanich, V. 2003. "A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis", *Proceedings of HLT-NAACL 2003*. Edmonton, Canada.

Teyssier, P. 1980. *História da Língua Portuguesa*. Lisboa: Sá da Costa.

Theobald, B. 2007. "Audiovisual Speech Synthesis", *International Congress of Phonetic Sciences 2007*, Saarbrücken, Germany.

Tokuda, K. 2004. "An HMM-Based Approach to Multilingual Speech Synthesis," *Shrikanth Narayanan, Abeer Alwan (Eds.), Text-to-Speech Synthesis: New Paradigms and Advances*. New Jersey: Prentice Hall.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. 1995. "An algorithm for speech parameter from continuous mixture HMM with dynamic features", *Proceedings of Eurospeech 95*, Madrid, Espanha. Pp. 757-760.

Trancoso, I. & Viana, M. C. 1997. "On the pronunciation mode of acronyms in several european languages", *Eurospeech 1997, 5th European Conference on Speech Communication and Technology*. Rhodes, Greece. pp. 573 – 576.

Trancoso, I.; Viana, M. C.; Silva, F.; Marques, G. and Oliveira, L.1994. "Rule-based vs. neural network based approaches to letter-to-phone conversion for Portuguese common and proper names", *Proceedings of ICSLP'94*, Yokohama, Japan. pp. 1767-1770.

Trancoso, I.; Céu Viana, M. 1995. "Issues in the Pronunciation of Proper Names: the experience of the Onomastica project," *Workshop on Integration of Language and Speech*, Moscow, Russia.

Turk, O., Schröder, M., Bozkurt, B., and Arslan L. M., 2005, "Voice Quality Interpolation for Emotional Text-To-Speech Synthesis", *Proceedings of Interspeech 2005*, Lisbon, Portugal. pp. 797-800.

Uto, Y.; Nankaku, Y., Toda, T.; Lee, A.; Tokuda, K. 2006. "Voice Conversion Based on Mixtures of Factor Analyzers", *Proceedings of Interspeech 2006*, Pittsburgh, USA. paper 2076-Thu1BuP.8

Veloso, J. 1999. *Na ponta da língua. Exercícios de Fonética do Português*. Porto: Granito Editores e Livreiros. ISBN: 972-8594-01-1.

Viana, M.C., d'Andrade, E. 1985. *CORSO I: um conversor de texto ortográfico em código fonético para o português*. Relatórios do grupo de fonética e fonologia n. 6, CLUL.

Villalva, A. 2003. "Formação de palavras: afixação", *Mateus, M. H. M. (coord.) Gramática da Língua Portuguesa*. Lisboa: Caminho.

Violaro, F. & Böeffard, O. 1998. "A Hybrid Model for Text-to-Speech Synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, no 5, pp. 426-434.

Waibel, A.; Bernardin, Wolfel, K.; M. 2007. "Computer-Supported Human-Human Multilingual Communication", *Proceedings of Interspeech 2007*. Antuérpia, Bélgica. pp. 14-21.

Weiss, C.; Paulo, S. Figueira, L.; Oliveira, L. 2007. "Blizzard Entry: Integrated Voice Building and Synthesis for Unit-Selection TTS", *Blizzard 2007*, ISCA.

Weiss, L. C.; Oliveira, L.; Paulo, S.; Mendes, C.; Figueira, L.; Vala, M.; Sequeira, P.; Paiva, A.; Vogt, T.; Andre, E. 2007. "ECIRCUS: Building Voices for Autonomous Speaking Agents", *6th Speech Synthesis Workshop*, ISCA.

Xavier, M. F. & Mateus, M. H. (orgs.) 1992. Dicionário de Termos Linguísticos da Associação Portuguesa de Linguística e do Instituto de Linguística Teórica e Computacional. Lisboa: Edições Cosmos. Vol. II.

Yang, Q.; Mertens, J. P.; Konings, N.; Heuvel, H. 2006. "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names", *Proceedings of LREC 2006*. Génova, Itália, pp. 287-292.

Yarowsky, D. 1996. "Homograph disambiguation in Text-to-Speech Synthesis", *Santen et al. (editors)Progress in Speech Synthesis.* pp. 159-174, New York: Springer.