Leitor de estrangeirismos para sistemas de conversão Texto-Fala em Português Europeu

Daniela Braga, Maria Aldina Marques
& Fernando Gil V. Resende Jr.
Microsoft Language Development Center, Universidade do Minho,
Universidade Federal do Rio de Janeiro

1. Resumo

A leitura de estrangeirismos, a par da desambiguação de homógrafos, constitui 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. Neste trabalho, 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 apresentados e discutidos.

2. Definição do problema e estado da arte

Apesar de se tratar de um problema com reduzida expressão nas línguas, a correcta leitura das palavras estrangeiras[1] aumenta substancialmente a qualidade perceptiva da

[1] 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)." O Dicionário da língua portuguesa contemporânea da Academia das Ciências de Lisboa (Casteleiro, 2001), conhecido pela sua modernidade no plano do tratamento dos estrangeirismos, define assim o conceito: "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 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

***Textos Seleccionados. XXIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2008, pp. 61-76**

síntese, tal como argumentam Llitjos et al. (2001): "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."

A leitura de estrangeirismos representa um dos problemas de mais difícil solução para a síntese da fala em Português, por vários motivos, nomeadamente: 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 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; 4) a escassez de trabalho linguístico sobre a integração de estrangeirismos na Língua Portuguesa e a ausência de inventários actualizados[2] e com transcrições fonéticas são aspectos que não facilitam o seu tratamento computacional; 5) tal como observado por Andrade & 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"; 6) os estrangeirismos apresentam diferentes graus de integração na língua de chegada, como descrito em Freitas *et al*. (2003)[3], sendo que na segunda fase de integração, por exemplo, se verifica a possibili-

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 (…)".

[2] Destacam-se três dicionários especializados de estrangeirismos para o PE: Costa (1990), Machado (1994) e Schmidt-Radefelt (1997), todos eles desactualizados e com palavras estrangeiras obsoletas e de uso duvidoso. A principal crítica que apontamos 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 site (http://www.iltec.pt/projectos/em_curso/portal.html, 20-01-2008), 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, 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 & Lopes (2003). No entanto, a sua busca torna-se bastante morosa, visto se tratar de dicionários que apenas dispõem de versão em papel.

[3] Segundo Freitas et al (2003), a integração dos estrangeirismos no PE atravessa três fases passando pelos seguintes fenómenos: 1) primeira fase: adaptações fonéticas e morfossintácticas imediatas, monossemia: manutenção do significado com o qual a palavra é importada, grafia da língua de origem e hesitação nos 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.

dade de formação de novas palavras segundo as regras morfológicas do Português (ex. <surfista> [s6rfiSt6]; <checkar> [SEkar]).

Apesar de representarem sempre uma pequena percentagem na língua, ou seja, cerca de 1,4% do léxico[4] e 1,3% em textos reais[5], os estrangeirismos são uma classe aberta, em permanente expansão na língua, designando uma grande panóplia de entidades, desde marcas, empresas, instituições, nomes próprios, topónimos, moedas, produtos, a termos técnicos e científicos, o que torna obrigatório o seu tratamento pelos sintetizadores de fala.

Entre as propostas de resolução da leitura de estrangeirismos, contam-se as técnicas por dicionário (Black *et al*., 1998), os modelos estatísticos CART-based (Llitios *et al*., 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. Não é nosso objectivo específico neste trabalho tratarmos os nomes próprios estrangeiros, embora muitos deles possam ser tratados sempre que forem abrangidos pelas regras.

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.

3. Leitor de estrangeirismos

Neste capítulo, apresenta-se um leitor de estrangeirismos integrado num sistema de conversão Texto-Fala em português europeu (PE). A concepção deste módulo, construído segundo regras linguísticas, assenta na identificação da língua de origem e passa por três fases (ver Figura 1): o pré-processamento, que inclui um dicionário com transcrição fonética de palavras que não provêm nem do inglês nem do francês; o

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

[5] Constitui-se um corpus de 7893 palavras, composto por textos do Expresso online de 25/09/2007 a 30/09/2007, extraídos a partir de várias secções: opinião, economia, desporto África, actualidade, emprego. Neste corpus, 102 palavras eram estrangeiras, representando assim 1,3% do corpus, sendo a maior percentagem encontrada nas secções de opinião e a menor nas secções de emprego.

identificador de galicismos e anglicismos (necessário por representarem o maior número de estrangeirismos em PE); 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.

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[6]), 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.

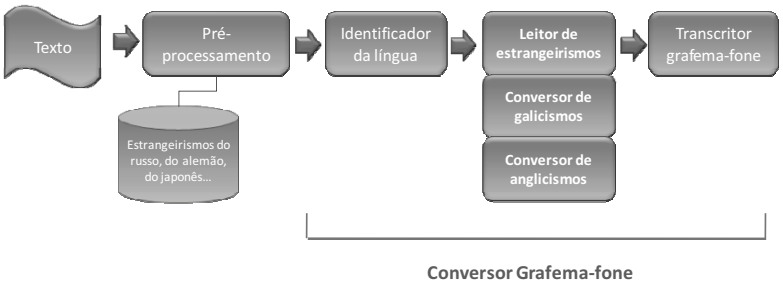

Figura 1: Arquitectura do leitor de estrangeirismos.

Outra dificuldade, já referida atrás, foi 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. Em seguida, passaremos a descrever as várias fases e funcionamento do leitor de estrangeirismos.

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

3.1. 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 PE. 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, como por exemplo <Apartheid, ayatollah, Beethoven, Dostoievski, kalashnikov, Volkswagen, etc.>. 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, contendo neste momento 30 palavras.

3.2. 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. Se alguma das condições do primeiro losango da Figura 2 se verificar, a palavra em análise é identificada como estrangeira.

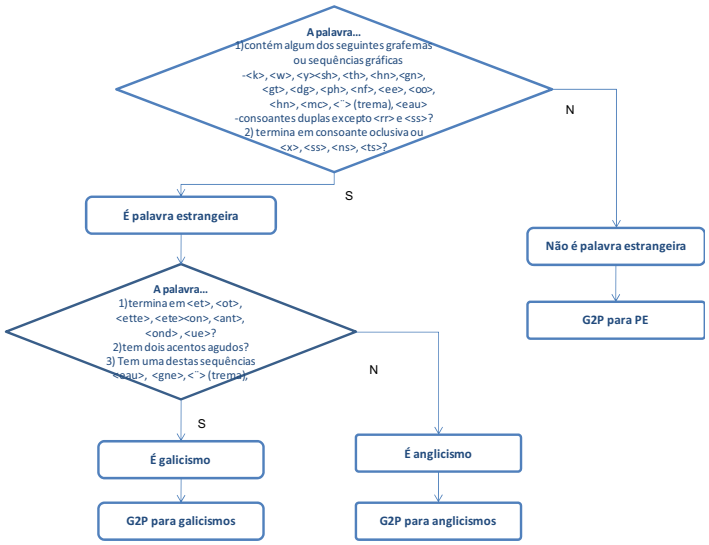

Figura 2: Algoritmo de identificação da língua.

Caso nenhuma das condições se verifique, a palavra passa para o conversor grafema-fone do PE (Braga *et al*., 2006; Braga *et al*., 2007). 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.

3.3. 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 (Simões *et al*., 2007), 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 constituída por 102 palavras. A lista completa destas palavras, bem como mais detalhes sobre o sistema, pode ver-se em Braga (2008).

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

• Consoantes nasais em posição pré-consonântica são associadas 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[7] (ex. coffee-break → co[f]ee-break);

• Paragoge de *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>)[8];

• 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).

Símbolo	Significado
...	Qualquer grafema
< x >	Grafema ou conjunto de grafemas <x>
,	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	Qualquer 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
<(W_bgn) x>	<x> está em início de palavra

Tabela 1: Símbolos e convenções de anotação usados no leitor de estrangeirismos.

Na Tabela 1, apresentam-se as convenções utilizadas para o desenho dos algoritmos de conversão grafema-fone das consoantes e vogais inglesas e francesas apresentados nas Tabelas 2, 3 e 4.

[7] Este fenómeno já tinha sido mencionado em Céu Viana *et al.* (1994).

[8] Este fenómeno também está atestado em Céu Viana *et al.* (1994).

#	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
#	**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][9]	Queens, Quick Silver
2	... < q >...	[k]	quark
#	**padrão gráfico de <r>**	**fone**	**Exemplo**
1	... < r r > ...	[R]	Ferry

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

2	...<(W_bgn) r>...	[R]	Rock
3	... < r >...	[r]	cross, Broadway
#	**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, aces
5	…<s><C_UV[10]>…	[S]	casting, desktop
6	…<s><SP, Pont>…	[Ø]	Excep. de pal. francesas[11]
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][12]	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

Tabela 2: Tabela de conversão das consoantes inglesas e francesas.

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 2. As regras começam pelas sequências gráficas mais raras, terminando com um *default*.

[10] Grafemas consonânticos não vozeados ou surdos: <p>, <t>, <k>, <q>, <c>.

[11] Lista de palavras cujo <s> final não se lê: <ménage à trois> [me.na.Za.trwa1], <collants> [kO.l6~1S], <croissants> [krwa.s6~1S].

[12] Em palavras de origem germânica, <w> articula-se [v]: <wagner, wagneriano, wálchia>.

#	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
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][13]	overhead, sweater[14]
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,
2	…<o u n><C>…	[a~w~]	country, background
3	…<o u><C>…	[aw]	check-out,
4	…<o w><C>…	[aw]	brownie[15]
5	…<o a>…	[O][16]	bodyboard, broadway
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

13 Excepção em: <leasing>,<sex-appeal>, <Shakespeare>, <features>, <strip-tease><ea> lê-se [i].

14 Excepção em: <breaveheart>, <ea> lê-se [a].

15 Excepção: <show> [So1w], <snowboard> [snow.bO1rd].

16 Excepção: lê-se [ow] em <ferryboat>, <roaming>.

10	…<o><r><SP, Pont, s>…	[6]	Communicator
11	…<o>…	[O][17]	golf, hip hop
#	**padrão gráfico de <u>**	**fone**	**Exemplo**
1	…<(W_bgn) C><u><C>…	[6][18]	bus, bluff, budget, blush,
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

Tabela 3: Tabela de conversão das vogais inglesas.

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 3, podem ver-se as regras de conversão grafema-fone das vogais para os anglicismos.

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 4, apresentamos as regras de conversão grafema-fone para as vogais dos galicismos.

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 de excepções 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 (ex. <áirbag>→[Er-b´E-g@]; <déadline>→[dE-d@-´laj-n@]; <ínternet>→[i~-tEr-´nE-t@]).

[17] Excepção: lê-se [6] em <motherboard> [m6.d6r.bO1rd].

[18] Excepção: lê-se [6] em <motherboard> [m6.d6r.bO1rd].

#	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
#	**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

Tabela 4: Tabela de conversão das vogais francesas.

A utilização directa do divisor silábico e do marcador de tonicidade automáticos (descritos em Braga *et al.*, 2007) sobre os estrangeirismos revelou-se pouco viável, dada a elevada taxa de erro: em 100 estrangeirismos escolhidos aleatoriamente do nosso corpus, o sistema errou 15,0% na separação silábica e 28,0% 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. Testes e discussão de resultados

Os vários módulos do leitor de estrangeirismos foram implementados em linguagem C++ e testados com corpora. Na Figura 3, pode ver-se a interface deste módulo integrado com outros módulos do *front-end* do sistema de conversão Texto-Fala para PE.

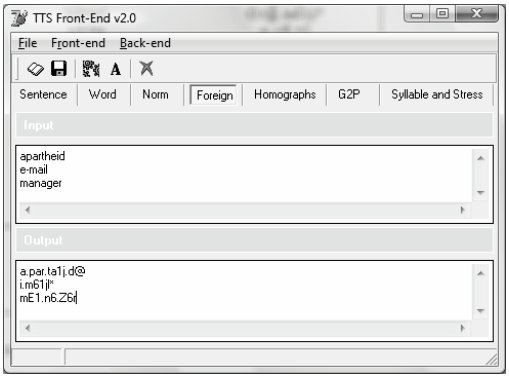

Figura 3: Interface do leitor de estrangeirismos.

Tipo de erro	#	%WER[19]	%PER[20]
<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**

Tabela 5: Resultados da avaliação do leitor de estrangeirismos.

Tendo por objectivo avaliar a importância deste módulo na arquitectura global do *front-end* do sintetizador, corremos um conjunto de 586 palavras estrangeiras e 3773 caracteres sem espaços no transcritor grafema-fone isoladamente (descrito em Braga *et al.*, 2007). Este corpus, diferente do que foi usado para formular os algoritmos, é composto por uma lista de palavras estrangeiras consideradas mais frequentes e inseridas num léxico fonético, essencialmente nomes de produtos, marcas, vocabulário comum, fornecido por uma empresa de software. A taxa de erro deste módulo por palavra testando palavras estrangeiras foi de 75,4%. Em seguida, corremos as mesmas 586 palavras no leitor de estrangeirismos, 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 5 ilustra os resultados deste teste. Os principais erros ocorrem na transcrição das vogais inglesas <e> (ex. <greatest hits> [grEtESt its], <sky news> [ski nEwS]), <a> (ex.

19 WER – word error rate (taxa de erro por palavra).

20 PER – Phone error rate (taxa de erro por fone).

<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]).

5. Conclusões

Neste trabalho, apresentou-se um módulo de leitura de estrangeirismos baseado em regras linguísticas integrado num sistema de conversão Texto-Fala em PE. Este trabalho teve como objectivo melhorar a taxa de acerto do conversor grafema-fone do sistema. Fez-se, numa primeira fase, o levantamento de estrangeirismos em Português. Em seguida, identificaram-se as suas origens. Finalmente, 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. A existência deste módulo representou uma melhoria de 63,45% ao nível da taxa de erro por palavra estrangeira. Dadas as diferentes origens dos nomes próprios estrangeiros, não foi nosso objectivo neste trabalho testar este módulo com esse tipo de palavras. Como trabalho futuro, pretendemos: 1) abordar mais extensivamente os nomes próprios estrangeiros; 2) comparar os resultados da abordagem por regras aqui enunciada com uma abordagem por métodos estatísticos ou *data-driven*; 3) tratar a marcação de acento tónico e a divisão silábica dos estrangeirismos de forma automática.

Referências

Andrade, Ana Rebello & Lopes, António Lavouras (2003) O tratamento dos estrangeirismos nas últimas edições do Dicionário da Língua Portuguesa da Porto Editora. In *Revista de Lexicografia da Universidada da Coruña*. vol. IX. A Coruña: Universidade da Coruña, pp. 7-28.

Bergström, Magnus & Reis, Neves (2007) *Prontuário ortográfico e guia da língua portuguesa*. Cruz Quebrada: Casa das Letras.

Black, Alan; Lenzo, Kevin and Pagel, Vincent (1998) Issues in building general letter to sound rules. In *Third ESCA/COCOSDA Workshop on Speech Synthesis*. Jenolan Caves House, Blue Mountains, Australia, pp.77-80.

Braga, Daniela & Resende Jr., Fernando Gil Vianna (2007) Módulos de Processamento de Texto Baseados em Regras para Sistemas de Conversão Texto-Fala em Português Europeu. In *XXI Encontro da Associação Portuguesa de Linguística*. Coimbra, 2-4 Outubro de 2006, pp.141-156.

Braga, Daniela (2008) *Algoritmos de Processamento de Linguagem Natural para Sistemas de Conversão Texto-Fala em Português*. Dissertação de Doutoramento. A Coruña: Universidade da Coruña.

Braga, Daniela, Coelho, Luís; Resende Jr., Fernando (2006) A Rule-Based Grapheme-to--Phone Converter for TTS Systems in European Portuguese. In *VI International Telecommunications Symposium (ITS2006).* Fortaleza-CE, Brasil, pp. 328-333.

Casteleiro, João Malaca (coord.) (2001) *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. 2 vols. Lisboa: Editorial Verbo.

Chen, Yining; You, Jiali; Chu, Min; Zhao, Yong; Wang, Jinlin (2006) Identifying language origin of person names with N-grams of different units. In *Proceedings of ICASSP2006*. Toulouse, France, pp. 729-732.

Costa, Francisco Alves (1990) *Dicionário de Estrangeirismos*. Lisboa: Editorial Domingos Barreira.

Estrela, Edite; Soares, Maria Almira; Leitão, Maria José (2004) *Saber escrever. Saber falar. Um guia completo para usar correctamente a língua portuguesa*. Lisboa: Dom Quixote.

Freitas, Tiago; Ramilo, Maria Celeste e Soalheiro, Elisabete (2003) O processo de integração dos estrangeirismos no Português Europeu. In Mendes & Freitas (orgs.) *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*. Lisboa, Portugal.

Garcia, Marie-Neige; Morel, Michel; Prudon, Romain; Véronis, Jean (2005) Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters. In *Proceedings of Interspeech 2005*. Lisbon, Portugal, pp. 1521-1524.

Grande Dicionário da Língua Portuguesa da Porto Editora (2004) Porto: Porto Editora.

Lavouras Lopes, António & Rebello d'Andrade, Ana (1997) Primeira fase da instalação do estrangeirismo. In *Actas do XIII Encontro da APL*. Lisboa: Colibri, pp. 77-89.

Llitjos, Ariadna Font & Black, Alan (2001) Knowledge of language origin improves pronunciation accuracy of proper names. In *Proceedings of Eurospeech 2001*. Alborg, Denmark, pp. 1919-1922.

Mareüil, Philippe Boula de; d'Alessandro, Christophe; Bailly, Gérard; Béchet, Frédéric;

Schmidt-Radefelt, Jurgen & Schurig, Dorothea (1997) *Dicionário dos Anglicismos e Germanismos na Língua Portuguesa*. Frankfurt am Main: Verlag Teo Ferrer de Mesquita.

Simões, Carla; Calado, António; Braga, Daniela; Teixeira, Carlos; Dias, Miguel (2007) 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, pp. 738-747.

Taylor, Paul (2005) Hidden Markov Models for Grapheme to Phoneme Conversion. In *Proceedings of Interspeech 2005*. Lisbon, Portugal, pp. 1973-1976.

Viana, Maria do Céu; Trancoso, Isabel; Silva, Fernando (1994) On the pronunciation of proper names and acronyms in European Portuguese. In *2nd Onomastica Research Colloquium,* London, December 1994.

Yang, Qian; Mertens, Jean-Pierre; Konings, Nanneke; Heuvel, Henk van den (2006)

Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme--to-phoneme (g2p) conversion of names. *In Proceedings of LREC 2006*. Génova, Itália, pp. 287-292.