CONVERSOR DE GRAFEMAS PARA FONES BASEADO EM REGRAS PARA PORTUGUÊS

Sara Candeias, Fernando Perdigão

IT – Instituto de Telecomunicações, pólo de Coimbra
Departamento de Engenharia Electrotécnica e de Computadores
Universidade de Coimbra
3030-290 Coimbra
email: {saracandeias, fp}@co.it.pt

I. Introdução

Esta apresentação tem por objectivo descrever um sistema de conversão automática de grafema para fone (GR2PH) para o português de Portugal. Para o desenvolvimento do sistema está a ser usado o corpus de unidades acentuais (palavras) em língua portuguesa, disponibilizado pela Universidade do Minho (proveniente da colaboração entre a Linguateca e o Projecto Natura). A avaliação do sistema fará uso do vocabulário da base de dados SPEECHDAT bem como de outros corpora de teste já usados por diversos investigadores a trabalhar neste domínio. A anotação fonética de corpora em língua portuguesa seria um interessante recurso linguístico a tornar público na Linguateca. Este recurso poderia tornar-se disponível, depois de avaliado e validado o sistema.

A crescente procura de soluções baseadas em produtos de tecnologia da fala tem sido uma motivação para o desenvolvimento de sistemas capazes de estabelecer um interface Homem--Máquina mais natural, como são exemplos as práticas subjacentes a áreas do ensino/aprendizagem do português e da linguística clínica.
A consciencialização da necessidade destes produtos mobilizou ao desenvolvimento de um sistema que convertesse, de forma automatizada, corpora grafados em corpora notados foneticamente (GR2PH).
O sistema de conversão de grafema para fone, do qual fazem parte os subsistemas 'divisor de sílabas' e 'marcador de tonicidade', é aquele para o qual o conhecimento linguístico contribui com um maior impacto e assume, no quadro geral do sistema final, uma maior eficiência. A estratégia adoptada para a estruturação do sistema GR2PH baseia-se em regras linguísticas cotejadas na estrutura da língua portuguesa. Para o desenvolvimento quer do sistema que transmuta grafema em fone, quer dos sistemas intermédios para divisão silábica e para marcação de sílaba tónica, foi usado o corpus de unidades acentuais (perto de 680000) em língua portuguesa, disponibilizado como recurso nascido da colaboração entre a Linguateca e o Projecto Natura. Na verdade, o acesso a este recurso resultou numa mais valia à performance do(s) sistema(s) que se pretendia(m) desenvolver, e os testes que foram sendo feitos, mesmo de forma faseada, mostraram-se basilares na fase de estruturação da arquitectura do(s) próprio(s) sistema(s), complementares e final.
Para o português europeu, alguns transdutores de grafema para fone baseados em regras surgem descritos na literatura existente [2] [3] [6] [7], [8]. Para a implementação das regras, em certos grupos, é reconhecida a importância da identificação da unidade silábica [2] [3] [6] [7]; noutros, é usada a informação da tonicidade da vogal [2] [3] [8]. A indispensabilidade de desenvolvermos um novo sistema de conversão GR2PH para o português de Portugal advém de factores como a escassa partilha dos algoritmos dos sistemas já implementados (dos quais poder-se-ia partir para um esforço de melhoramento do sistema) e dos resultados dos testes de desempenho provenientes de estudo comparativos. Este artigo apresenta uma tessitura alternativa de regras linguísticas a serem aplicadas no GR2PH para o português de Portugal, aliando a pertinência da informação linguística de regras de silabificação e de marcação de tonicidade. Resultando o sistema final da configuração de dois subsistemas perspectivados em regras inerentes à língua, o esforço do investimento tem por objectivo a viabilidade de um conversor capaz de uma eficácia que torne dispensável o recurso a dicionários de excepções. A arquitectura deste sistema GR2PH é resultado da complementaridade da

aplicação do conhecimento linguístico e da ciência de engenharia, parceria esta que se traduz num diálogo necessário a uma execução que se pretende optimizada e eficaz.

II. Arquitectura do Sistema de conversão Gr2Ph

É característica deste sistema GR2PH o recurso a sistemas intermédios, como o de separação da unidade acentual (UA, palavras) em sílabas e o de marcação de sílaba tónica (e consequente delimitação de sílaba(s) pré-tónica(s) e de sílaba(s) pós-tónica(s)). A vantagem desta abordagem explica-se pelo facto de ela permitir resolver a quase totalidade de casos de escolha fonética que não seria a acertada se resultasse apenas da inserção dos fones (nomeadamente vocálicos) considerados a partir de inventários fonéticos não diferenciados, isto é, não ponderados nem silabicamente nem atendendo à tonicidade em âmbito contextual de UA.

Todas as regras foram implementadas inicialmente em Matlab e foram testadas no vocabulário da base de dados SPEECHDAT [5] e no corpus de unidades acentuais disponibilizado pela Linguateca/Projecto Natura.

Esta segunda parte apresenta as especificidades dos subsistemas de divisão silábica, de marcação de tonicidade e do transcritor, de forma a se ter uma visão global do sistema geral de conversão GR2PH. Na Tabela 1 figuram as convenções usadas nas regras para implementação.

Tabela 1: Convenções usadas nas regras para implementação.

convenções	significado
C	consoante
V	vogal
.	divisor de sílaba
'	marcador de tonicidade
#	fronteira final de UA
|	ou

a. Subsistema de divisão silábica

A estrutura deste subsistema assenta a) num modelo de regras de divisão de base ortográfica, b) na consideração de vogal como núcleo de sílaba e c) na consideração de alguns dígrafos como grafema singular (‘ch’, ‘ss’, ‘lh’, ‘gu’+‘i’|‘e’, ‘qu’+ ‘i’|‘e’, etc.). O algoritmo do ‘divisor de sílabas’ reproduz uma busca feita por padrões de até 5 grafemas, resultando em 18 possíveis encontros de sequências que formam sílaba em português de Portugal (Tabela 2). As regras foram distribuídas por dois grandes grupos para cada padrão de sequência de grafemas, isto é, considerando se na sílaba da UA a analisar é pertinente a informação dos 4 caracteres ou de mais que os 4 caracteres da sequência. Nesta repartição, surgem explícitas regras que apresentam um tipo repetido subsequente da iteração de sequências, como é exemplo a sequência VV presente nos padrões CCVV, CVVC, CVV e VVC. Na Tabela 4, a título de exemplificação de procedimentos, surgem descritas regras para o padrão CVVC.

b. Subsistema de marcação de tonicidade

Na estruturação deste subsistema, toda a unidade (palavra) foi considerada acentual (UA) e, por isso, não foram admitidos segmentos desprovidos de tonicidade [4]. O algoritmo de marcação da sílaba tónica funciona com regras instituídas a partir da divisão silábica. Admitiu-se o acento tónico como o acento da UA (o acento principal), pelo que, nesta estrutura, não se considerou pertinente marcar os acentos secundários. Na tabela 4 figuram regras de marcação de sílaba tónica.

Tabela 2: Lista dos padrões de sequências de grafemas a formar sílaba em português de Portugal.

sequência	exemplo
CCVCC	*trans*.cre.ver
CCVVC	grãos

CVCCC	*tungs*.té.ni.o
CCCV	*stre*.sse
CCVC	*trás*
CCVV	*grão*
CVCC	*subs*.cre.ver
CVVC	*mães*
VCVC	*achar*
VVC	*aus*.cul.tar
VCC	*abs*.tra.ir
CCV	a.*cre*
CVV	*pai*
CVC	a.*cam*.par
VC	*ac*.tu.*ar*
CV	*ca*.*sa*
VV	*eu*
V	*á*.gua

Tabela 3: Ilustração de algoritmo de divisão silábica para o padrão de grafemas CVVC.

	C	**V**	**V**	**C**		**grafema final da UA**	**grupo silábico**	**exemplo**
sequência		a|e|o|u a|e|o	i u	l|r|m|s|j	**V**		**CVV.C**	*pau*.lada, *mou*.ro, *tei*.ma, *lou*.sa, *bei*.jo
		ã|õ ã	e o	≠s|#				*mãe*.zinha, *mão*, ta.*lão*
	g|q	u	a|o		**V**			*quo*.ciente, *gua*.rida, *qua*.se, *qua*.lidade
		a|e|o|u a|e|o	i u	l|z	**#**		**CV.VC**	*pa*.ul, *ra*.iz
		a|e|o|u a|e|o	i u	r|m	**C|#**			*ca*.ir, *ru*.im, *co*.imbra
		a|e|o|u	i	nh	**V**			*ba*.inha, ta.*bu*.inha, *mo*.inho
		a|e|o|u a|e	i u	n	**C**			*re*.incide, tran.*se*.unte
		a|e|o|u a|e|o	i u	s	**C|#**		**CVVC.**	*cais*, *faus*.to, a.*zuis*, *bois*
		ã|õ ã	e o	s				*mãos*, *pães*
	g|q	u	a|o	l|n|r	**C|#**			*qual*, *qual*,quer, *guar*.da, *quan*.do
	por defeito						**CV.VC**	*be*.ata, *fi*.os

c. Subsistema de Gr2Ph

Para a anotação fonética, seguimos o alfabeto SAMPA para o português [1], sem o recurso a extensões como seria o caso das «oclusivas orais sonoras» «fricatizadas», traço que advém da posição em início de sílaba e intervocálica. Ainda que se tenha em vista a construção de um sistema de síntese futuro, o que leva a ter em conta, entre outros aspectos, a natureza particular de cada som em contexto de co-articulação e/ou de sandhi, o facto deste mapeamento da transmutação grafema--fone ir ser adicionado a um modelo acústico baseado em trifones, anula a necessidade de uma anotação fonética mais estreita. Com este mesmo princípio, não foram consideradas como «semiconsonânticas» 'j' e 'w' as unidades vocálicas grafadas 'i'(ou 'e') e 'u' (ou 'o') dos ditongos ditos crescentes (presentes em relóg*io* e em ár*ea*, em s*ua*ve e em nód*oa*). O algoritmo da conversão do grafema em fone funciona a partir das sílabas com 'marcação de tonicidade'. Isto é, a partir de um contexto-base, resultam casos de grafemas admitidos à conversão em fones que consideram a pertinência de informação da a) posição de tonicidade e da b) posição no âmbito da sílaba (na qual é pertinente o comportamento fonético dados os grafemas vizinhos). Na Tabela 5 são exemplificados algoritmos de conversão do grafema 'o' para os fones [o~], [w~], [o], [O] e [u], que resultam da atenção aos parâmetros descritos.

Tabela 4: Algoritmo de marcação de sílaba tónica.

	regra	**marcador de tonicidade**	**exemplo**
1.	Se na sílaba existirem vogais com acento gráfico	sílaba em questão	a.'*ná*.li.se
2.	Se na sílaba não existirem vogais sem acento gráfico		
2.1.	Se a UA tiver 1 sílaba	sílaba em questão	'*voz*
2.2.	Se a UA tiver ≥ 2 sílabas		
2.2.1.	Se for a última sílaba da UA com estrutura de: a|e|i|o|u + l|r|z i|u + ∅|s i + m	sílaba em questão	pa.'*ul* ra.'*iz* *ca.*'ir an.'*dou*, ca.pi.'*tais* pe.'*ru*, pe.'*rus* ru.'*im*
2.2.2.	por defeito	penúltima sílaba	a.na.'*li*.se

Tabela 5: Ilustração de algoritmo de conversão do grafema 'o' para fones.

fone	**Posição de tonicidade**	**Posição silábica**	exemplos
o~		+ m|n (mesma sílaba)	'*o*m.bro → *o*~bru; p*o*n.tu.'al → p*o*~tual
w~		ã + (mesma sílaba)	'cã*o* → k6~*w*~; cã*o*.'zi.nho → k6~*w*~ziJu
o	tónica	+ nh (sílaba seguinte)	ri.'s*o*.nho → riz*o*Ju
O	tónica	+ x (sílaba seguinte)	pa.ra.'d*o*.xo → p6r6d*O*ksu
o	tónica	+ i (mesma sílaba)	'*o*i.to → *o*jtu
o	tónica	+ r (mesma sílaba e final de UA)	pa.ssa.'d*o*r → p6s6d*o*r
O	tónica	+ r (mesma sílaba)	'c*o*r.ta → k*O*rt6
o	tónica	+ a (sílaba seguinte e final de UA)	't*o*.da → t*o*d6
O	tónica	por defeito	'*o*.de → *O*d@; 'c*o*.rre → k*O*R@
O	átona	(inicial de UA) + r	*O*r.ga.'ni.za → *O*rg6niz6
u	átona	+ r (mesma sílaba)	c*o*r.'tar → k*u*rtar
O	átona	(inicial de UA)	*o*.'ní.ri.co → *O*niriku
u	átona	o (sílaba anterior) +	co.*o*. pe.ra.'ção → ku*u*p@r6s6~w~
u	átona	(final de UA)	'fi.lh*o* → fiL*u*
O	átona	+ c|p (mesma sílaba)	*o*c.'ta.vio → *O*taviu; *o*p.'ção → *O*ps6~w~
u	átona	por defeito	p*o*.'ção → p*u*s6~w~

A análise e verificação de muitas regras foi conseguida por análise exaustiva ao corpus de UAs disponibilizado pela Universidade do Minho. Transcrições ou pronunciações alternativas não são consideradas neste sistema, como é o caso de homógrafos heterófonos.

III. Conclusão e Trabalho futuro

Até esta fase, a forma gráfica convertida automatizadamente em forma fonética foi avaliada com referência à anotação manual. Dispomos apenas do vocabulário associado à base de dados SPEECHDAT como material de teste, embora a avaliação com este corpus não esteja ainda concluída, especialmente devida à discordância encontrada na conversão das semiconsoantes dos ditongos crescentes. Uma forma alternativa de fazer a avaliação do sistema consiste em comparar os resultados de vários sistemas de conversão (pelo menos um é de domínio público [2]), contando e analisando as diferenças encontradas. Como trabalho futuro, pretendemos construir uma

aplicação on-line de conversão de grafemas para fones bem como de um corpus anotado foneticamente.

IV. Referências

[1] (SAMPA) Speech Assessment Methods Phonetic Alphabet. http://www.phon.ucl.ac.uk/home/sampa/portug.htm (2008/07/09)

[2] ALMEIDA, José João; SIMÕES, Alberto Manuel (2001). "Text to Speech: "a rewriting system approach"", Congreso de la SEPLN. 17, Jaén, 2001. [S.l. : s.n.], [c. 2001]. http://hdl.handle.net/1822/633 (2008/07/15)

[3] 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", Lobo, M. & Coutinho, M. A. (Orgs), *XXI Encontro da Associação Portuguesa de Linguística*. Coimbra, 2-4 Outubro de 2006.

[4] CANDEIAS, Sara (2007). "Vocalismo dos 'clíticos'", *Sistema fonológico da Beira Interior e algumas considerações sintáctico-semântica*. Tese de doutoramento em linguística, parte II, 4. Universidade de Aveiro: Departamento de Línguas e Culturas.

[5] (SPEECHDAT) Portuguese SpeechDat(II) FDB-4000, European Language Resources Association (1998). http://www.elda.org/catalogue/en/speech/S0092.html

[6] TEIXEIRA, António; OLIVEIRA, Catarina; MOUTINHO, Lurdes (2006). "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.

[7] TEIXEIRA, João Paulo; GOUVEIA, Paulo, FREITAS, Diamantino (2000). "Divisão Silábica Automática do Texto Escrito e Falado", *Proceedings of PROPOR'2000*. Atibaia, SP. Brasil.

[8] VIANA, Maria do Céu; D' ANDRADE, Ernesto (1985). *CORSO I: um conversor de texto ortográfico em código fonético para o português*. Relatórios do grupo de fonética e fonologia n. 6, CLUL.