gɾɐfɔnə: uma ferramenta de/com recurso à informação linguística

Sara Candeias, Arlindo Veiga & Fernando Perdigão

Instituto de Telecomunicações – Coimbra, Portugal,
Depto. de Eng. Eletrotécnica e de Computadores – Universidade de Coimbra, Portugal

Abstract

This paper describes the architecture of a new grapheme to phoneme converter in European Portuguese. Named as gɾɐfɔnə, this converter is presented as a resource available to the linguistic community. The converter is available at http://www.co.it.pt/~labfala/g2p/ and is able to generate pronunciation dictionaries. Linguistic information, namely on the phonological context and stressed vowel was incorporated in statistical algorithms, deeply improving the performance of the converter.

For this work a pronunciation dictionary with over 40,000 words derived from the CETEMPúblico corpus was constructed and is also available. Orthographic norms, with and without the Orthographic Agreement of 1990 were considered as well. The performance of the converter, which can be verified by the users exploring the tool available on the web, along with the panorama of on-going authors' studies, confirms the crucial role of the linguistic information in technological solutions development based on the Portuguese language. Keywords/ Palavras-chave: Grapheme to phoneme converter, pronunciation, phonology, stressed vowel /Conversão garfema-fonema, pronunciação; fonologia, vogal tónica

1. Introdução

O processamento da língua portuguesa em geral, e, na sua vertente falada em particular, tem vindo a sofrer um crescente interesse por parte da comunidade científica nos últimos anos. Este interesse provém das reais necessidades sentidas pelo mundo atual, envolto numa explosão de novas tecnologias, as quais trazem novas formas de acesso à informação e ao conhecimento. As tecnologias da fala são, assim, de grande interesse, não só como forma de tornar a comunicação homem-máquina o mais natural possível, como também pela possibilidade de tornar mais facilitada a comunicação entre os humanos (Llisterri, 2002; Candeias, 2011). Aplicações desenvolvidas para cidadãos mais idosos ou com necessidades especiais e para o ensino da língua são, em nosso entender, emergentes, e implicam um grande investimento ao nível de construção de recursos linguísticos e de ferramentas básicas. O estudo aqui apresentado acompanha essa tendência e descreve a arquitetura do gɾɐfɔnə, um conversor de grafema para fonema em língua portuguesa, vertente europeia (PE). O desempenho deste conversor

Textos Selecionados, XXVIII Encontro Nacional da Associação Portuguesa de Linguística, **Coimbra,APL, 2013, ppp. 189- 203, ISBN: 978-989-97440-2-8**

foi francamente melhorado com a incorporação de informação linguística (contextos fonológicos e regras de acentuação), o que vem confirmar, a par do trabalho que os autores têm vindo a desenvolver, o papel crucial do conhecimento linguístico para o desenvolvimento de sistemas tecnológicos de fala. Juntamente com a ferramenta conversor, disponível na web em http://www.co.it.pt/~labfala/g2p/, a qual permite a que a uma lista de palavras ou léxico, se faça corresponder a respetiva transcrição fonológica, disponibiliza-se igualmente um dicionário de pronunciação com mais de 40000 vocábulos derivados do corpus CETEMPúblico (SANTOS, 2001). Acrescenta-se que as normas ortográficas, sem e com o Acordo Ortográfico de 1990, foram consideradas no desenvolvimento do conversor. A ferramenta de conversão funciona com ambas as normas.

Comummente conhecido por G2P (do inglês *Grapheme to Phoneme* [1]), o mapeamento entre grafemas e fonemas tem por objetivo converter um texto escrito numa sequência de símbolos que representam os sons da fala de uma determinada língua, de uma forma inequívoca. Ainda que a investigação na área, em PE, seja já sólida e madura, a taxa de desempenho dos sistemas de G2P, como facilmente se comprova pelos erros de conversão que persistem nos sistemas existentes no mercado, ainda não atingiu o nível de desempenho desejado. No sentido de encontrar uma solução para os problemas da conversão de G2P para o PE, são várias as abordagens que têm vindo a ser propostas, de entre as quais destacamos as que derivam de regras linguísticas (Braga, 2006; Oliveira, 1992; Teixeira, 2004); de regras inferidas a partir dos dados (Teixeira, 2006); por máquinas de estados finitos (Caseiro, 2002; Oliveira, 2004); por máxima entropia (Barros, 2006); baseadas em redes neuronais (Trancoso *et al.*, 1994); e por CARTs - *Classification and Regression Trees* (Oliveira *et al.*, 2001). A abordagem por modelos probabilísticos (Demberg, 2007; Bisani, 2008) é a que tem sido referenciada por contraste à que resulta de regras linguísticas e tem sido essencialmente proposta para línguas cujo sistema ortográfico está mais distante da estrutura fonológica. Na verdade, para línguas como o inglês ou francês, as quais não apresentam uma clara correspondência entre o grafema e fone(ma), a solução encontrada para o problema da conversão automática pode basear-se no pressuposto de que é possível prever a pronunciação de um vocábulo, por analogia, a partir de exemplos suficientes de sequências identificativas dessa correspondência. Ainda que vantajosa, por não implicar uma revisão constante, e muitas vezes morosa, das regras, em especial quando surgem sequências distintas das regularidades previamente admitidas, a conversão G2P sustentada apenas por modelos probabilísticos também não capta contextos suficientes de forma a espelhar a estrutura fonológica da língua. Em relação ao PE, por exemplo, um modelo auxiliado por regras linguísticas justificaria a presença de uma acentuação secundária em <v**a**gamente> /vˌ**a**gɐmˈẽtə/[2], na medida em que o conhecimento dado apenas pela estatística de sequência de grafemas associaria, erroneamente, ao vocábulo, a pronúncia /v**ɐ**gɐmˈẽtə/, em analogia à sequência <v**a**ga->, de pronunciação /v**ɐ**/,

[1] Por vezes também designado L2S: *Letter to Sound*.
[2] Faz-se uso do Alfabeto SAMPA (Wells, 1997) (cf. Tabela 1).

posição átona em relação ao acento primário, presente em <v**a**gar> /v**ɐ**gˈaɾ/, <v**a**gabundo> /v**ɐ**gɐbˈũdu/ ou em <v**a**garoso> /v**ɐ**gɐɾˈozu/, por exemplo.

Assim como as línguas românicas na sua generalidade, o PE apresenta uma regularidade fonética e fonológica bem como uma ortografia de base fonológica, características que justificam o sucesso da inclusão, em algoritmos de sistemas de conversão G2P, de regras para marcação de sílaba tónica (Candeias, 2008; Braga, 2006; (Almeida, 2001; Teixeira, 1998).

O sistema de G2P que aqui se apresenta teve na base da sua construção uma abordagem híbrida na qual concorrem modelos estatísticos/probabilísticos imbuídos de informação linguística. Essa informação linguística, resultante de regras da estrutura fonológica do PE, é aqui exposta, chamando-se a atenção para o que delas resultou em termos de melhoramento do desempenho do conversor. O desempenho do conversor, que pode ser testado pelo utilizador no uso da ferramenta disponível na web (http://www.co.it.pt/~labfala/g2p/), aliada ao panorama de estudos em andamento pelos autores, vem assim mostrar o papel crucial da informação linguística no desenvolvimento de sistemas de processamento do português com vista à apresentação de soluções tecnológicas baseadas na língua. Para além do mais, para o desenvolvimento do ɡrɐfɔnə, foi construído um dicionário de pronunciação com mais de 40000 vocábulos derivados do corpus CETEMPúblico, um recurso que também se encontra disponível em SPL (2011).

A apresentação deste trabalho encontra-se estruturada da seguinte forma: na secção seguinte apresenta-se brevemente o modelo híbrido usado. A secção 2 descreve a informação linguística introduzida no modelo probabilístico adotado. Os resultados do desempenho do sistema de conversão G2P são dados na secção 5, à qual se seguem as conclusões.

2. Modelo híbrido

O modelo adotado para a tarefa de converter grafemas em fonemas faz uso de um modelo estatístico/probabilístico, no qual se integra informação linguística em forma de regras. A classificação de híbrido dá conta dessa natureza. De forma muito breve, apresentamos nesta secção o modelo estatístico/probabilístico. A informação linguística imbuída nesse modelo será descrita na secção seguinte. Para se obter mais pormenores técnicos sobre a operacionalidade do modelo, a qual se julga além do universo do presente documento, leia-se Veiga (2012).

Usando uma abordagem estatística/probabilística, a determinação da sequência ótima de fonemas, dada a sequência de grafemas, é condição necessária para a tarefa de conversão entre grafemas e fonemas. Qualquer abordagem estatística adotada na tarefa de conversão requer a existência de um dicionário fonológico, necessário para estimar as probabilidades dos padrões encontrados, e a maioria das abordagens requer ainda um

algoritmo que permita o alinhamento entre grafemas e fonemas. A favor de um alinhamento unívoco entre grafema e fonema, concorre a particularidade do PE apresentar uma ortografia de base fonológica como é o caso de muitas das consoantes, como o <t>, em <linguís**t**ica> o qual é diretamente convertido no fonema /t/. Contudo, situações que derivam de casos como o dos grafemas <n> e <u> no mesmo exemplo (<li**ngu**ística>) revelam que a correspondência entre grafemas e fonemas pode estar dependente de fatores contextuais. Outras situações ainda, como as que resultam da associação entre os grafemas <e> e <o> e os respetivos fonemas, chamam a atenção para a dependência do estatuto morfológico, algumas das vezes com interdependência sintática. Vejam-se os casos: <s**e**lo> (nome) → /sˈ**e**lu/ e <s**e**lo> (verbo) → /sˈ**ɛ**lu/ e <**o**lho> (nome) → /ˈ**o**ʎu/, <**o**lho> (verbo) → /ˈ**ɔ**ʎu/. Existem situações em que um único grafema pode originar vários fonemas, assim como existem situações em que vários grafemas podem originar um único fonema (como <g> e <j> → /ʒ/ em exemplos como <**g**esto> → /**ʒ**ˈɛʃtu/ e <**j**eito> → /**ʒ**ˈɐitu/[3], conversão esta igualmente dependente de contexto, vocálico ao caso).

Todas as abordagens estatísticas/probabilísticas deparam-se com problemas decorrentes do alinhamento. A solução encontrada para, durante o processo de treino do sistema, segmentar e alinhar grafemas e fonemas com igual número de segmentos, nem sempre é trivial ou única e depende da forma como os algoritmos de alinhamento associam os grafemas aos fonemas de um dado vocábulo. Seguindo a classificação de alinhadores proposta por Jiampojamarn (2007), para desenvolver o ɡrɐfɔnə, usámos o alinhador de "um-para-um", na vertente de "1-01", como se descreve em Veiga (2012). À entidade que resulta desse alinhamento, a qual é composta pela associação de um segmento de grafemas a um segmento de fonemas, damos o nome de grafonema (na linha de Bisani (2002). Visualize-se 6 grafonemas do vocábulo <compõem> na Figura 1. Os modelos usados para estimar a probabilidade de um grafonema também se encontram descritos em (Veiga, 2012).

Grafemas	*c*	*om*	*p*	*õ*	*e*	*m*
Fonemas	k	o ~	p	o ~ i ~	ɐ ~	i ~

Figura 1: Exemplos de grafonemas. Cada entidade grafonema encontra-se entre parênteses retos.

A opção pelo modelo de "1-01", no qual cada grafema dá origem a zero ou a um fonema, foi desde logo tomada pela verificação de que, em apenas 7 casos do PE, um grafema pode dar origem a mais do que um fonema. Esses casos, com respetivos contextos de ocorrência são indicados de seguida:

(1) extrair → /ɐiʃtrɐˈiɾ/ (<e> → /ɐi/)

(2) extra → /ˈɐiʃtrɐ/ (<e> → /ˈɐi/)

[3] Sobre a notação de glides cf. opções descritas em 3.2.

(3) têm → /tˈɐ̃ĩɐ̃ĩ/ (<ê> → /ˈɐ̃ĩɐ̃/)

(4) põem → /pˈõĩɐ̃ĩ/ (<õ> → /ˈõĩ/)

(5) axila → /aksˈilɐ/ (<x> → /ks/)

(6) caem → /kˈaiɐ̃ĩ/ (<a> → /ˈai/)

(7) constroem → /kõʃtɾˈɔiɐ̃ĩ/ (<o> → /ˈɔi/)

Os grafemas suscetíveis de serem apagados são <u|c|m|n|p|r|s|h|z>. Os grafemas <c> e <p> são apagados apenas quando se converte o vocabulário grafado na forma prévia à aplicação do Acordo Ortográfico de 1990 (AO).

Para a integração das regras linguísticas no modelo estatístico/probabilístico como ele foi assumido ("1-01"), teve que se operar uma transformação na sequência de grafemas, introduzindo-se novos símbolos. Eles são símbolos uni caráter convencionados e estão descritos em Veiga (2011).

3. Informação linguística

Nas subsecções seguintes, descrevem-se constrições linguísticas do PE pertinentes à tarefa de converter o grafema em fonema, as quais foram adicionadas ao módulo de G2P. Algoritmos baseados em regras fonológicas para a acentuação vocálica, reconhecendo o núcleo de sílaba tónica de cada vocábulo, e para a identificação da correspondência exata entre um grafema e respetivo fonema, de acordo com o contexto, foram propostos.

3.1. Vocabulário

Para que um dicionário de pronunciação fosse gerado pelo sistema de conversão de grafema em fonema, foi necessário, numa primeira fase, ter disponível como base de trabalho uma listagem de vocábulos atuais e representativos do PE. O material utilizado para esse fim foi o corpus CETEMPúblico (Santos, 2001), o qual contém 180 milhões de palavras[4], provenientes de uma coleção de extratos do jornal Público de entre os anos 1991 e 1998.

O processo de criação dessa listagem consistiu em tomar todas as cadeias de carateres anotadas como palavras, obedecendo simultaneamente aos seguintes critérios:

i. começar com um grafema do alfabeto português: <a-z>, <A-Z>, <á-ú>, <Á-Ú>;

[4] Por palavras entendem-se, aqui, todos os átomos do corpus que contêm, pelo menos, um grafema ou dígito. Neste trabalho, adotamos como sinónimos os termos ‘palavra’, ‘vocábulo’ e ‘unidade acentual’.

ii. não conter dígitos;
iii. não apresentar todos os grafemas em maiúscula (caso de siglas);
iv. não conter o caráter '.' (caso de URLs);
v. terminar com um grafema do alfabeto português ou com '-';
vi. o lema correspondente não conter o caráter '=' (caso de nomes compostos).

A partir do resultado obtido, formou-se uma lista de cerca de 50k vocábulos, correspondentes a uma contagem de ocorrências no corpus de mais do que 70 vezes. Os nomes próprios não foram considerados. Sendo arbitrária, a consideração deste número de ocorrência para a configuração do vocabulário de base deveu-se especialmente ao facto de anular a possibilidade de se estarem a incluir erros tipográficos muitas vezes frequentes nas edições de imprensa. Por fim, foram retirados quer vocábulos estrangeiros quer estrangeirismos, usando-se critérios automáticos e validação manual. Todos os vocábulos que apresentavam grafemas ou sequências grafemáticas que não fazem parte do sistema do PE, tais como <k>, <w> e <y>; <sh> e <pp>; e <b>, <d> ou <p> em posição final de vocábulo foram excluídos. Alguns destes dados serviram para a constituição de um dicionário de pronunciação de cerca de 1300 estrangeirismos. Como resultado final deste processo, uma lista de cerca de 40k vocábulos foi gerada. É esta lista que corresponde ao vocabulário de referência tomado para este trabalho, cuja referência é "voc_CETEMP_40k". Para se constituir uma listagem adicional com vocábulos grafados de acordo com as normas anteriores ao AO, submeteu-se o vocabulário pré-AO ("voc_CETEMP_40k") à ferramenta Lince (Lince). Dos 41586 vocábulos utilizados no vocabulário pré-AO, cerca de 2% sofreu alterações de grafia, tais como a eliminação das consoantes mudas (<c> e <p>) e dos hífenes e a alteração de acentuação gráfica. Perante a possibilidade de duas grafias coexistirem, este novo vocabulário apresenta pares de vocábulos ditos ‘parónimos’, tais como <conceptual> e <concetual> ou <desconectar> e <desconetar>. O vocabulário pós-AO é constituído, assim, por 41598 vocábulos, e é referenciado como "voc_CETEMP_40k_ao".

3.2. Transcrição fonológica

Neste momento, uma dilucidação acerca da opção pela transcrição fonológica, e não fonética, é devida. É comummente aceite pela comunidade linguística que a fonética diz respeito às propriedades físicas e articulatórias de todos os sons que ocorrem na produção linguística, cabendo à fonologia o estudo da função de cada som pronunciado numa dada língua, a qual permite ao falante distinguir significados. É igualmente aceite que qualquer opção metodológica no que à análise da fala diz respeito, liga, inevitavelmente, as duas faces do binómio, uma vez que lida tanto com a relação que existe entre as unidades e a sua pertinência na língua falada (i.e., os fonemas) como com a realidade física que resulta na pronunciação dessas mesmas unidades (i.e., os fones e alofones) (cf. definições dos termos em Crystal, 2001). Tem sido frequente a alternância, muitas vezes não claramente justificada, entre os termos fone e fonema nos vários estudos efetuados no âmbito do G2P (a título de exemplo, a unidade fone é a adotada em Caseiro (2002), enquanto que Barros (2006) apresenta o

fonema como o resultado da conversão do grafema). Neste estudo, considerámos trabalhar ao nível do fonema, uma vez que o procedimento de conversão adotado admite valências do contexto mais ou menos alargado no âmbito da unidade acentual (vulgo palavra), considerando a unidade para a qual o grafema é convertido como uma escolha significativa por entre todas as outras unidades que o sistema de língua coloca ao dispor. Assim, aceitamos a unidade fonológica, ou fonema, como uma classe à qual pode corresponder um fone ou um feixe de realizações alofónicas disponíveis no PE (acolhendo-se, assim, a inserção de pronúncias alternativas – processo que está já em andamento pelos autores deste texto). A transcrição fonológica resultante corresponde ao PE que admitimos como padronizado e não representa qualquer arquifonema ou neutralização de oposições. A transcrição é registada entre barras oblíquas, e fizemos uso, neste processo, do alfabeto SAMPA (Wells, 1997) (cf. Tabela 1).

Relativamente à transcrição fonológica do vocabulário de referência, o processo ocorreu de forma iterativa, e que passamos a descrever.

Em primeiro lugar, foi feito um modelo estatístico/probabilístico, conforme referido supra, tendo por base o dicionário de pronunciações da base de dados SpeechDat (SpeechDat) com cerca de 15k vocábulos. Para a constituição do dicionário foram retirados os estrangeirismos e foram feitas algumas correções de pronunciação. Na processo foram convencionadas algumas particularidades, as quais tiveram por base essencialmente critérios acústicos de uniformização. São elas:

i. os símbolos representativos das glides /j/ e /w/ foram notados como as vogais correspondentes /i/ e /u/;
ii. não se distinguiu a lateral velarizada da lateral, ainda que sistemas reconhecidos de anotação para o português, como o usado na SpeechDat, admitam a presença de /ɬ/ (/5/ em X-SAMPA) e de /l/.
iii. admitiu-se a necessidade de inclusão do iode, nomeadamente para nos aproximarmos, o mais possível, do PE padronizado (por exemplo, os vocábulos iniciados por <ex-> são transcritos como /ɐiʃ/, como em <extra> /ˈɐiʃtrɐ/ em contexto de tonicidade) (ver exemplos na secção 2);
iv. os símbolos SAMPA adotados (cf. Tabela 1) resultam da ponderação sobre representatividade do PE falado (Wells, 1997).

Neste ponto, julga-se com pertinência, dilucida-se o seguinte: uma observação atenta dos alfabetos fonéticos SAMPA para o Português (SAMPA-PT) e X-SAMPA dá conta de alguma indefinição de regularidade, exemplificada na atribuição de mais do que um símbolo para o mesmo som. Na verdade, o símbolo [r] no SAMPA-PT parece ter como correspondente no X-SAMPA o símbolo [4] (IPA: [ɾ]), simbolizando o [r] no X-SAMPA a vibrante alveolar múltipla (IPA: [r]).

Seguiu-se então um processo moroso de confirmação e correção manual das transcrições obtidas automaticamente. O passo seguinte consistiu em comparar as transcrições do dicionário com as transcrições geradas por um sintetizador de fala

comercial. Esta comparação permitiu-nos confiar no nosso resultado já que, maioritariamente, as transcrições coincidiram. As transcrições que diferiram foram analisadas individualmente e corrigidas quando necessário, no sentido da representatividade do PE. Deste processo resultou o dicionário de transcrição fonológica que referenciamos como "dic_CETEMP_40k", o qual também se encontra disponível em (SPL, 2011).

SAMPA	Grafemas possíveis	Exemplos
6	a, e, â, ê	c**a**ma, s**e**nha, c**â**mara, am**ê**ijoa
a	a, á, à	p**á**, p**a**la, **à**
@	e	d**e**
e	e, ê	v**ê**, d**e**do
E	e, é	p**é**, p**e**le
i	i, í, e, y	v**i**, a**í**, r**e**al, henr**y**
o	o, ô, ou	**o**co, av**ô**, l**ou**co
O	o, ó	p**ó**, p**o**te
u	u, ú, o, w	t**u**, ti**o**, at**o**, ba**ú**, ki**w**i
6~	ã, an, ân, am, e, âm, é, a	v**ã**, br**an**co, **ân**cora, c**am**po, t**e**m, l**âm**pada, al**é**m, i**a**m
e~	ên, en, em, êm	p**en**te, ag**ên**cia, **em**pate, **êm**bolo
i~	i, in, im, ím, ín, m, n, e	mu**i**to, tr**in**car, s**im**, **ím**pio, **ín**timo, home**m**, be**n**s, põ**e**
o~	õ, ôn, ôm, on, om	p**õe**, c**ôn**sul, c**ôm**puto, p**on**te, p**om**bo
u~	u, ún, un, um, úm, o, m	m**u**ito, an**ún**cio, **un**s, at**um**, c**úm**plice, vã**o**, ia**m**
b	b	**b**e**b**er
d	d	**d**a**d**o
g	g	**g**ato, **gu**elra
p	p	**p**ato
t	t	**t**oca
k	q, c, k	**q**uando, **c**asa, **k**iwi
f	f	**f**é
s	s, ç, x, c, ss	**s**ol, ca**ç**a, trou**x**e, **c**éu, a**ss**im
S	ch, s, z, x	**ch**ave, pá**s**, pa**z**, **x**á
v	v	**v**ida
z	z, s, x	ca**s**a, **z**ebra, e**x**emplo
Z	j, g, s, z, x	**j**á, **g**ira, de**s**viar, e**x**-líder
l	l	**l**âmpada
L	lh	ve**lh**o
r	r	ca**r**o
R	r, rr	ca**rr**o, **r**ato
m	m	**m**ão
n	n	**n**ada
J	nh	se**nh**a

Tabela 1: Símbolos SAMPA associados a grafemas possíveis com vocábulos exemplificativos.

Ao longo do desenvolvimento do conversor, o dicionário tem sofrido um processo constante de revisão e de correção. Apesar de admitirmos a presença de alguns erros de transcrição, estamos confiantes na sua precisão, pelo que acreditamos que o dicionário

"dic_CETEMP_40k" constitui uma base de trabalho interessante para estudos na língua portuguesa, em especial na área da fonética e da fonologia. Do processo que acompanhou este estudo, resultou igualmente um dicionário de pronunciação de cerca de 1300 estrangeirismos. Este dicionário de estrangeirismos será incorporado no sistema de conversão, como tabela de exceções.

3.3. Vogal tónica

A marcação da vogal tónica ($V_{tónica}$), núcleo de sílaba, e consequente alternância entre vogais tónicas e átonas, tem impacto ao nível do conversor, como têm vindo a provar trabalhos prévios de conversão de G2P, tais como Candeias (2008), Braga (2006), Caseiro (2002), Almeida (2001), Teixeira (1998), quer para a implementação de regras de transmutação do grafema em fone(ma), quer para a modulação de índices prosódicos (em especial se a informação for alargada à sílaba tónica).

Seguindo os pressupostos teóricos de Mateus (2000) e Andrade (1985) na consideração da pertinência da marcação da $V_{tónica}$ (identificada com o símbolo SAMPA ") e não da respetiva unidade silábica, o processo de identificação de $V_{tónica}$ que adotamos não seguiu, tanto quanto nos é dado a perceber, o perfil que tem sido usual noutros trabalhos. Assim, o algoritmo de marcação de $V_{tónica}$ foi definido e imbuído no modelo estatístico/probabilístico. O algoritmo começa por atender ao contexto vocálico grafemático de cada vocábulo, e, se alguma vogal (<*V*>) recebe um acento gráfico, essa <*V*> é identificada como $V_{tónica}$ (cf. Tabela 2, regra 1). Caso não apresente graficamente qualquer marca de tonicidade, ignora-se o <s> final e analisa a última <*V*> nos vocábulos terminados em <i>, <u>, <im>, <um>, <in>, <un> ou em consoantes, <*C*>, exceto <m> e <n> (cf. regra 2, Tabela 2). Os restantes vocábulos (sem grafemas acentuados), recebem indicação de $V_{tónica}$ em posição paroxítona (Tabela 2, regra 3)

A aplicação das regras descritas são suficientes para não marcar tonicidade nos vocábulos com uma única <*V*> não acentuada graficamente, como são os casos:

i. das preposições <com>, <de>, , <sem> e das contrações <do(s)>, <no(s)>;
ii. dos pronomes pessoais oblíquos <me>, <te>, <se>, <nos>, <vos>, <lhe(s)>, <o(s)> e <a(s)>, <lo(s)>, <no(s)>, <vo(s)> e das contrações <mo(s)>, <to(s)>, <lho(s)>; do pronome relativo <que>; e das conjunções <e>, <nem>, <que>, <se>, as quais se agregam frequentemente a um grupo de força acentual no âmbito do sintagma prosódico.

Existem 32 verbos que contêm sequências de <qu> ou <gu> onde o <u> é pronunciado e que pode aparecer na posição tónica (exs. <aguar> → <ag**u**e>, <aguir> → <arg**u**e>, <adequar> → <adeq**u**e>, <delinquir> → <delinq**u**e>).

	Regras	Exemplos
1	**Se** vocábulo apresenta alguma <*V*> acentuada graficamente, e**ntão** <*V*> → <$V_{tónica}$> [5].	auxí**l**io, an**á**lise, avaliaç**ã**o, **à**s, t**ú**nel
2	**Se** vocábulo não apresenta acento gráfico e ignorando o <s> final, se terminar em <i>, <u>, <im>, <um>, <in>, <un> ou <C> exceto <m> e <n>, e**ntão** a última <*V*> → <$V_{tónica}$>.	cant**a**r, emit**i**r, dev**e**r, can**a**l, pap**e**l, fun**i**l, cet**i**m, telef**a**x, dupl**e**x, cab**a**z, delf**i**m, bot**i**ns, par**i**s, alg**u**m, com**u**ns, jes**u**s
3	**Se** vocábulo não apresenta acento gráfico e ignorando o <s> final, terminar em <a>, <e> ou <o>, seguido ou não de <m|n>, e**ntão** a penúltima <*V*> → <$V_{tónica}$>	c**a**rta, d**a**nça, d**a**nçam, cont**e**nte(s), h**o**mem, h**o**mens, est**u**do(s)
4	**Se** em 2 e 3 a <$V_{tónica}$> for <i> ou <u> e se for precedida de uma <*V*> diferente de <i> e < $V_{tónica}$> e não no contexto <qu> ou <gu>, e**ntão** essa outra <*V*> → <$V_{tónica}$>.	p**a**i(s), r**e**i(s), l**e**u, m**a**u(s), decid**i**u, c**a**ixa(s),ad**e**us, p**e**ixe, p**a**uta(s)
5	**Se** em 4, a <*V*> <i> ou <u> é seguida de <m + C|#> e <n + *C*> ou <l + *C*|#> *C* exepto <h> e <l> ou <r + *C*|#> *C* exceto <r> ou <z + *C*|#> *C* exceto <z>, e**ntão** <i> ou <u> → <$V_{tónica}$>.	Coimbra, amendo**i**m, ra**i**nha, Ra**u**l, ca**i**rmos, ra**i**z

Tabela 2: Algoritmo para a marcação de acento tónico nas vogais (<*V*>).

3.5. Contextos frequentes

A descodificação da associação entre grafema e fonema sem ambiguidade foi também auxiliada pela indicação de regras simples que atendem ao contexto grafemático. A título de exemplo, a determinação da sequência grafemática <al+*C*> resulta na notação de <a> em /a/ (em <almoçar> → /almusˈaɾ/); a definição de <*V*+s+*V*> resulta na notação de <s> em /z/ (em <casa> → /kˈazɐ/). Foram ainda definidas outras regras para o <s> e para os grafemas <r>, <z>, <c>, <g> e <x>, inseridos em contextos mais restritos. Considerando um contexto mais alargado, na sequência grafemática <muit>, as <V_{orais}> <u> e <i> passam a /V_{nasais}/.

4. Resultados

Todas as experiências foram baseadas no dicionário de pronunciação de 41586 vocábulos da língua portuguesa, descrito na subsecção 3.1. Aplicando diferentes formas de pré-processamento ao dicionário base, foram criados vários outros dicionários, disponíveis em (SPL, 2011), entre eles:

i. Dicionário base (dic_CETEMP_40k);

[5] Palavras como órfão(s), órfã(s), órgão(s), sótão(s), ímã(s), embora apresentem mais do que um acento gráfico, têm apenas uma sílaba tónica (em posição paroxítona) mas são marcadas com duas vogais tónicas.

ii. Dicionário com acentuação: presença da marcação da vogal tónica em cada pronunciação (dic_CETEMP_40k_acentuado)

Para testar o modelo estatístico, cada um destes dicionários foi particionado em 5 dicionários de treino e 5 dicionários de teste, de forma rotativa. O dicionário inicial foi dividido em 5 partes, cada uma com 20% dos vocábulos (8317), escolhidos de forma aleatória. Os vocábulos foram mutuamente exclusivos em cada uma das 5 partes. Cada uma das partes deu origem a um dicionário de teste e os restantes 4 partes (33269 vocábulos) a um dicionário de treino. A rotação das partes deu origem a 5 ciclos de treino e teste dos modelos estatísticos para validação cruzada. Os resultados indicados correspondem à média dos 5 resultados parciais.

O desempenho do sistema de conversão de G2P é expresso em duas taxas médias de erros de conversão verificados nos dicionários de teste: taxa média de erro de fonemas (PER – "phoneme error rate") (Figura 2) e taxa média de erro de vocábulos (WER – "word error rate") (Figura 3).

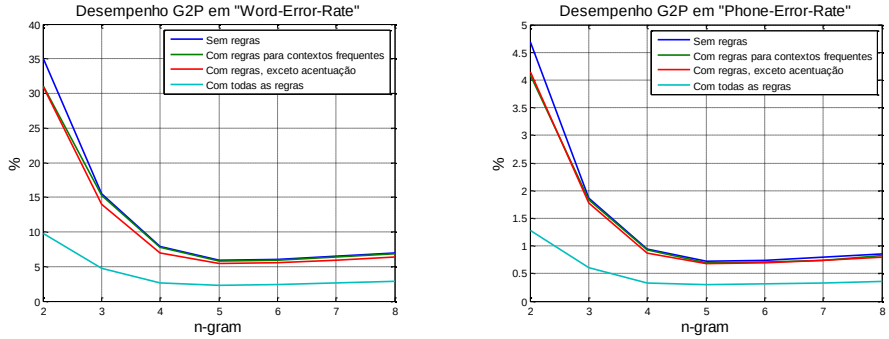

Figura 2 (à esquerda): Taxas de erro de palavras em função da inclusão de regras fonológicas e do comprimento do "*n*-grama". Por "*n*-grama" entendem-se grafonemas com contexto de comprimento, onde *n* é finito. Ao caso, "5-grama" é o ponto ótimo. Figura 3 (à direita): Taxas de erro de fonemas em função do comprimento do "*n*-grama" e da inclusão de regras fonológicas.

Os gráficos da Figura 2 e da Figura 3 ilustram o contributo de cada etapa de pré-processamento fonológico no desempenho do sistema de conversão, apresentando as percentagens da taxa de erro de conversão de vocábulos. Como se pode observar, a informação linguística é relevante, sendo a marcação da vogal tónica é o processamento que mais contribui para o melhoramento do desempenho do sistema gɾɐfɔnə.

5. Conclusões

Neste documento fez-se a descrição da arquitetura de um novo conversor de grafema em fonema para o português europeu: o gɾɐfɔnə. Este conversor apresenta-se

como um recurso disponível à comunidade linguística, acessível em http://www.co.it.pt/~labfala/g2p/. Ele permite gerar dicionários de pronunciação, que consistem numa lista de palavras ou léxico, a que corresponde a respetiva transcrição fonológica. É possível optar por transcrição SAMPA ou IPA (International Phonetic Alphabet), assim como evidenciar a vogal tónica.

Na génese do desenvolvimento do ɡrɐfɔnə, sequências de grafonemas foram modeladas através de um algoritmo de alinhamento entre grafemas e fonemas, nas quais foram consideradas informações advenientes do contexto fonológico da língua portuguesa (nomeadamente acentuação tónica e a vizinhança fonético-fonológica). Todas estas informações linguísticas foram testadas individualmente, tendo-se verificado que a inclusão de informação sobre a tonicidade da vogal foi decisiva para o aumento do desempenho do conversor. As normas ortográficas, sem e com o Acordo Ortográfico de 1990, forma consideradas.

Para este trabalho foi igualmente construído um dicionário de pronunciação com mais de 40000 vocábulos oriundos do corpus CETEMPúblico, do qual derivaram outros dicionários. Todos os dicionários estão disponíveis em SPL (2011).

O desempenho do ɡrɐfɔnə, o qual pode ser verificado pelo utilizador no uso da ferramenta disponível na web, aliada ao panorama de estudos em andamento pelos autores, vem confirmar o papel crucial da informação linguística no desenvolvimento de sistemas de processamento do português com vista à apresentação de soluções tecnológicas baseadas na língua.

O dicionário de estrangeirismos e o dicionário de múltipla pronunciação de homógrafos serão incluídos no sistema, a breve prazo. A pronunciação de adjetivos, de verbos e de nomes flexionados encontra-se em estudo avançado, também com o objetivo de vir a integrar o sistema ɡrɐfɔnə.

6. Agradecimentos

Este trabalho recebeu o apoio de fundos nacionais através de FCT – Fundação para a Ciência e Tecnologia (bolsa SFRH/BPD/36584/2007, projeto PTDC/CLE-LIN/11 2411/2009, e financiamento multianual PEst-OE/EEI/LA0008/2011 - Instituto de Telecomunicações).

Referências

Almeida, J. J. & Simões, A. (2001) Text to Speech – A Rewriting System Approach. *Procesamiento del Lenguaje Natural*, 27, pp. 247–255.

Andrade, E. & Viana, M. C. (1985). *Curso I - Um Conversor de Texto Ortográfico em Código Fonético para o Português*. Lisboa: Technical report, CLUL-INIC.

Barros, M. J. & Weiss, C. (2006). Maximum Entropy Motivated Grapheme-To-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech. Zaragoza(Spain): IV Jornadas en Tecnologías del Habla, pp. 177-182.

Bisani, M. & Ney, H. (2002). Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion. Denver(USA): Proc. 7th International Conference on Spoken Language Processing (ICSLP'02), pp. 105–108.

Bisani, M. & Ney, H. (2008) Joint-Sequence Models for Grapheme-To-Phoneme Conversion. Speech Communication, vol. 50 (5), pp. 434–451.

Braga, D., Coelho, L. & Resende Jr. F. (2006) A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese. Fortaleza-CE(Brazil): VI Int. Telecommunications Symposium, pp. 328-333.

Candeias, S. & Perdigão, F. (2011) Integração Linguística em Sistemas de Conversão de Grafema para Fone(ma). In Luís,Ana R. (ed.). *Estudos de Linguística*. Coimbra: Imprensa da Universidade de Coimbra,.vol 1.

Candeias, S. & Perdigão, F. (2008) Conversor de Grafemas para Fones Baseado em Regras para Português. In Costa, L.; Santos, D.; Cardoso, N. (Eds.). *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*. Lisboa: Linguateca, cap. 14.

Caseiro, D. A. & Trancoso, I. (2002) Grapheme-to-Phone Using Finite-State Transducers. USA: Pro. 2002 IEEE Workshop on Speech Synthesis.

Crystal, D. (2001) *A Dictionary of Linguistics and Phonetics*. Blackwell, Oxford.

Demberg, V., Schmid, H. & Möhler, G. (2007) Phonological Constraints and Morphological Preprocessing for Grapheme-to-phoneme Conversion. Prague(Czech Republic): Proc. 45th Annual Meeting of the Association for Computational Linguistics (ACL-07), pp. 96-103.

Jiampojamarn, S., Kondrak, G. & Sherif, T. (2007) Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. Rochester(New York): HLT-NAACL, pp. 372-379.

Lince - Conversor para a Nova Ortografia.
http://www.portaldalinguaportuguesa.org/lince.php

Llisterri, J. & Martí, M. A. (2002) *Tratamiento del Lenguaje Natural*. Barcelona: Edicions de la Universitat de Barcelona, S.L. Unipersonal.

Mateus, Maria Helena & d'Andrade Ernesto (2000) The Phonology of Portuguese, Oxford University Press.

Ney, Hermann, Essen, Ute & Kneser, Reinhard. (1994) On Structuring Probabilistic Dependences in Stochastic Language Modelling. *Computer Speech and Language*, vol. 8 (1), pp. 1-38.

Oliveira, C., Moutinho, L. & Teixeira, A. (2004) Um Novo Sistema de Conversão Grafema-Fone para PE Baseado em Transdutores. Brasil: Actas II Congresso Int. Fonética e Fonologia.

Oliveira, L., Viana, M. C., Mata, A. I. & Trancoso, I. (2001) *Progress Report of Project Dixi+: A Portuguese Text-to-Speech Synthesizer for Alternative and Augmentative Communication*. FCT: Technical Report.

Oliveira, L., Viana, M. C. & Trancoso; I. (1992) A Rule-Based Text-to-Speech System for Portuguese. San Francisco(USA): Proc. ICASSP'92,

Santos, D. & Rocha, P. (2001) Evaluating CETEMPúblico, a Free Resource for Portuguese. Toulouse(France): Proc. 39th Annual Meeting of the Association for Computational Linguistics, pp.442-449.

SpeechDat. - Databases for the Creation of Voice Driven Teleservices, http://www.speechdat.org/SpeechDat.html

SPL (2011) - Material disponibilizado no âmbito deste artigo, http://lsi.co.it.pt/spl/resources.htm

Teixeira, A., Oliveira, C. & Moutinho, L. (2006) On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion. Rio de Janeiro(Brazil): Proc. PROPOR'2006, pp. 212-215.

Teixeira, J. P. & Freitas, D. (1998) MULTIVOX- Conversor Texto-Fala para Português. Porto Alegre(Brasil): Proc. PROPOR'98.

Teixeira, J. P. (2004) *A Prosody Model to TTS Systems*. PhD Thesis, Faculdade de Engenharia da Universidade do Porto.

Trancoso, I., Viana, M. C., Silva, F., Marques, G. & Oliveira, L. (1994) Rule-based vs. Neural Network Based Approaches to Letter-to-Phone Conversion for Portuguese Common and Proper Names. Yokohama(Japan): Proc. ICSLP'94, pp. 1767-1770.

Veiga, A., Candeias, S. & Perdigão, F. (2011) Conversão de Grafemas para Fonemas em Português Europeu – Abordagem Híbrida com Modelos Probabilísticos e Regras

Fonológicas. J. J. Almeida, A. Simões, X. Guinovart (eds.). *LinguaMÁTICA, Revista para o Processamento Automático das Línguas Ibéricas*, vol 3, nº2: 39-51.

Veiga, A., Candeias, S. & Perdigão, F. (2012) Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment. *Journal of the Brazilian Computer Society*, 88 (JBCS). Springer.

Wells, J. C. (1997) SAMPA Computer Readable Phonetic Alphabet. Gibbon, D., Moore, R. and Winski, R. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter, part IV.