M.Céu VIANA (CLUL),
Isabel M. TRANCOSO (INESC/IST),
Fernando M.SILVA (INESC/IST),
Gonçalo MARQUES (INESC),
Ernesto d´ANDRADE (FLUL/CLUL),
Luís C. OLIVEIRA (INESC/IST)

Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu

Introdução

O desempenho dos sistemas de regras de conversão grafema fone para nomes próprios é, em geral, muito inferior ao observado para o léxico comum. Este facto nada tem de surpreendente, uma vez que a maior parte dos sistemas de regras existentes foram optimizados justamente para o léxico comum e que só raramente têm sido contempladas peculiaridades da pronúncia dos nomes próprios. Ele afecta, no entanto, o desempenho global dos sintetizadores de fala e justifica, por si só, um estudo mais cuidado. Existem, no entanto, outros factores que também podem estar na origem das diferenças de desempenho observadas e que é necessário ter em consideração. Os nomes de origem estrangeira, por exemplo, contêm muitas vezes sequências de segmentos que não obedecem às restrições fonotácticas da língua do país de acolhimento e o grau de adaptação da sua pronúncia à estrutura sonora dessa língua pode ser muito variável. Por outro lado, a ortografia dos nomes próprios nativos pode ser bastante conservadora, apresentando sequências de letras que não são contempladas e que, naturalmente, se prestam a interpretações fonéticas incorrectas. Os nomes de empresas levantam também sérios problemas uma vez que nem sempre obedecem às regras gerais de ortografia e de pronúncia.

Esta comunicação foca algumas destas questões para o Português Europeu. Nas secções seguintes, descrever-se-ão brevemente os corpora de nomes próprios e de léxico comum utilizados neste estudo comparativo, assim como alguns dados

estatísticos deles extraídos. Comparar-se-ão, em seguida, duas metodologias diferentes utilizadas para a conversão grafema-fone: sistema de regras e rede neuronal multi-camada. Terminar-se-á com um estudo, necessariamente breve, sobre a pronúncia e constituição dos nomes de empresas e serviços públicos.

Uma grande parte deste trabalho enquadra-se dentro do projecto nacional *BDFALA* (Programa Lusitânia) e do projecto europeu *Onomastica* (Programa LRE*).* Na sua realização foram, no entanto, utilizadas ferramentas de trabalho desenvolvidas no âmbito do projecto *DIXI* (convénio INESC / CLUL) e no âmbito da actividade do grupo de redes neuronais do INESC. De entre estas, destacam-se o corpus de frequência PF_Fone, um subconjunto de programas que asseguram a conversão grafema fone, o alinhamento automático das formas ortográficas e suas respectivas transcrições fonéticas [14,19,20] e, ainda, o módulo de redes neuronais.

1. Descrição dos corpora

Os corpora utilizados para o estudo dos nomes próprios foram construídos com base no material fornecido pela operadora de telecomunicações nacional participante no projecto Onomástica (TLP). A partir dos cerca de 100.000 nomes de pessoas, ruas, localidades e empresas do corpus original (palavras isoladas), foram constituídos vários subconjuntos:

- Nomes_Fone1: subcorpus constituído pelos 20.000 nomes mais frequentes das listas telefónicas de Lisboa e Porto.

- Nomes_Fone2: subcorpus de cerca de 15.000 nomes constituído a partir do anterior, excluindo estrangeirismos, erros de grafia, siglas e acrónimos.

- Nomes_Fone3: subcorpus de cerca de 12.000 nomes extraídos do anterior, excluindo nomes de empresas e designações de serviços públicos que são também formas do léxico comum.

- Acro_Fone: subcorpus de cerca de 21.000 nomes de empresas e designações de serviços públicos, incluindo acrónimos e siglas presentes na base de dados dos TLP e, ainda, um conjunto de siglas extraídas a partir de um corpus de jornais nacionais.

A estrutura de qualquer destes subcorpora é semelhante: cada entrada contém uma forma única, a indicação da sua frequência de ocorrência e a sua transcrição fonética. As transcrições foram geradas automaticamente com o sistema de regras desenvolvido no âmbito do programa DIXI e, depois, processadas manualmente[1] para correcção dos valores fonéticos atribuídos aos segmentos, das marcas de

[1] Uma grande parte das correcções manuais foi realizada por duas bolseiras do INESC: Ermelinda Gonçalinho e Catarina Moraes.

acento e da localização das fronteiras de sílaba. Todas as entradas foram ainda classificadas, também manualmente, em função da sua língua de origem e de diferentes categorias: *Nome de baptismo, Apelido, Nome de rua, Nome de edifício, Nome de lugar ou região, Nome comum, Nome de Companhia ou empresa, Acrónimo*[2] *ou Sigla*.

De modo a efectuar um estudo comparativo relativamente ao léxico comum foi também utilizado o corpus PF_Fone [14,20], construído a partir do corpus Português Fundamental [11], recolhido pelo CLUL. O corpus PF_Fone contém cerca de 26.000 formas de citação e formas flexionadas, com a respectiva frequência e transcrição fonética, esta última corrigida manualmente.

Dos cerca de 100.000 nomes (isolados) do corpus original de nomes próprios, sensivelmente metade constituem ocorrências únicas. Ordenando este corpus por ordem decrescente de frequência, os primeiros 13.000 nomes ocorrem mais de 10 vezes e os primeiros 2.700 mais de 100 vezes. Com base no subcorpus de frequência superior a 10, consegue-se uma cobertura de 88% dos nomes completos existentes na lista de Lisboa, 91% dos existentes na lista do Porto e 84% dos existentes nas listas do resto do país. Em termos de cobertura de nomes individuais, as percentagens são ainda maiores: 96% e 93%, respectivamente para as listas das duas cidades e do resto do país. A cobertura nacional dos subcorpora utilizados é, portanto, bastante significativa. No quadro seguinte, apresentam-se os valores de percentagem calculados para as várias categorias no subcorpus Nomes_Fone1, excluindo sucessivamente do cálculo os que pertencem a uma das categorias anteriores. (Exemplo: a quarta linha mostra a percentagem de entradas que são classificadas como apelidos, mas não como primeiros nomes ou topónimos). A última linha corresponde a 4% de nomes estrangeiros e a 2% de erros de grafia, cujas formas não foram classificadas quanto à categoria.

CATEGORIA	%
Primeiro nome	16
Topónimo	17
Apelido	28
Companhia (comum)	17
Companhia (acrónimo ou sigla)	16
Formas não classificadas	6

Quadro 1 - Distribuição por categorias no subcorpus Nomes_Fone1

[2] O termo *Acrónimo* é utilizado aqui como designação geral para nomes de empresas ou serviços públicos que não são formas do léxico comum nem nomes de baptismo ou apelidos, mas que podem corresponder a diferentes tipos de combinações da totalidade ou da parte de todos eles.

A maior parte das entradas pertence a múltiplas categorias. O quadro 2, em que foram ignorados os nomes de empresas e as formas não classificadas (Nomes_Fone3), mostra os resultados cruzados para as três primeiras categorias do quadro 1 e, ainda, o cruzamento destas com formas do léxico comum. Qualquer nome que pertença à lista do Português Fundamental ou a um conjunto de formas de citação e formas flexionadas gerado a partir de um dicionário com cerca de 86.000 entradas foi considerado como forma do léxico comum.

Prim. Nom	Apelido	Topónimo	Léx. comum	%
+	-	+	+	0,2
+	-	+	-	0,2
-	-	+	-	1,3
+	-	-	+	1,7
+	+	+	-	1,8
+	+	+	+	2,2
+	+	-	+	2,5
-	-	+	+	2,5
+	+	-	-	6,7
-	+	+	-	8,4
+	-	-	-	10,4
-	+	+	+	14,9
-	+	-	+	20,8
-	+	-	-	26,2

Quadro 2 - Ocorrências de formas (%) com base numa classificação cruzada de categorias, não tendo em consideração a frequência de ocorrência das formas.

A classificação foi feita automaticamente, utilizando um simples processo de verificação, e não é, naturalmente, exaustiva. Apesar disso, pode verificar-se que cerca de 45% dos nomes próprios analisados são formas do léxico comum. É interessante notar também que cerca de 84% dos nomes próprios analisados ocorrem como apelidos e que mais de um terço destes últimos são também topónimos. Repare-se ainda que, embora virtualmente qualquer primeiro nome possa ocorrer como apelido, este facto apenas se verifica para cerca de metade das formas desta classe.

Com base nos corpora PF_Fone e Nomes_Fone3, foi efectuada uma análise comparativa da distribuição de grafemas e fones no léxico comum e nos nomes próprios. Nem a análise pesada em frequência, isto é, tendo em conta o número de ocorrências de cada entrada, nem a simples, mostraram diferenças muito significativas como se pode observar nos quadros 3 e 4 que apresentam a distribuições de grafemas e fones, respectivamente.

GRAFEMA	PF	PF(FR)	NOM	NOM(FR)
a	13,6	12,3	15,9	14,0
ã	0,4	1,2	0,6	0,6
á	0,5	0,9	0,3	0,3
à	< 0,1	0,1	0,0	0,0
â	0,1	< 0,1	0,1	0,1
e	9,9	12,3	8,3	9,3
é	0,2	1,1	0,4	0,8
ê	0,1	0,2	0,1	0,1
i	7,9	5,9	8,9	8,6
í	0,4	0,2	0,3	0,4
o	8,4	10,1	9,5	9,8
ó	0,2	0,2	0,2	0,6
ô	< 0,1	< 0,1	< 0,1	0,0
õ	0,2	0,01	0,1	0,1
u	2,9	4,8	2,7	3,4
ú	0,1	0,1	0,1	0,1
b	1,3	0,9	1,9	1,1
c	4,5	3,0	4,2	2,9
ç	0,6	0,3	0,4	0,5
d	4,3	4,0	3,1	2,9
f	1,2	0,9	1,1	1,3
g	1,5	1,0	2,1	1,6
h	1,4	1,4	2,3	0,9
j	0,3	0,3	0,5	1,4
l	2,9	2,4	5,4	4,7
m	4,2	5,4	3,2	4,0
n	5,3	4,9	5,9	6,2
p	2,6	2,8	1,9	1,5
q	0,4	1,9	0,4	0,5
r	8,0	5,9	8,9	9,2
s	7,8	8,1	5,3	7,6
t	5,2	4,7	3,7	3,7
v	1,8	1,4	1,5	1,8
x	0,4	0,2	0,2	0,2
z	0,6	0,6	0,6	0,2
-	1,3	0,4	0,0	0,0

QUADRO 3- Distribuição de grafemas no léxico comum (PF) e nos nomes próprios (NOM) não pesada e pesada (FR).

FONE	PF	PF(FR)	NOM.	NOM(FR)
ũ	0,2	0,8	0,1	0,1
ʎ	0,3	0,3	0,6	0,3
ɬ	0,7	0,8	1,6	2,4
ɲ	0,7	0,6	1,1	0,3
e	0,7	2,0	1,0	0,6
ĩ	0,7	0,7	0,7	0,8
ʃ̃	0,7	1,3	0,2	0,1
w	0,8	1,0	0,8	1,8
ʒ	0,9	0,7	1,1	1,8
o	0,9	1,8	1,1	0,8
õ	0,9	0,7	0,4	0,6
ɛ	1,1	2,3	1,7	2,3
w̃	1,1	1,6	0,6	0,6
g	1,2	0,8	1,8	1,4
ʀ	1,2	0,5	1,7	1,5
ɔ	1,2	1,1	1,4	1,8
f	1,3	1,0	1,2	1,4
b	1,4	1,0	2,1	1,2
z	1,4	1,2	1,3	1,5
n	1,5	2,3	3,0	2,9
ẽ	1,7	1,2	0,5	0,5
v	2,0	1,5	1,6	1,9
l	2,1	1,5	3,6	2,3
j	2,4	2,2	3,7	4,0
ɛ̃	2,5	3,3	1,8	2,5
p	2,8	3,0	2,1	1,5
m	2,9	3,4	2,8	3,9
k	3,6	4,4	3,5	2,7
s	3,6	3,5	2,7	3,0
a	4,6	4,2	3,8	3,4
d	4,7	4,3	3,4	3,0
i	5,5	6,4	3,8	4,5
t	5,6	5,2	3,9	4,0
ʃ	5,9	5,3	4,2	5,3
i	6,3	4,7	6,0	5,2
ɾ	7,0	5,8	7,1	7,7
u	7,6	7,6	7,9	7,9
ɐ	10,2	10,1	14,3	12,4

QUADRO 4 - Distribuição de fones no léxico comum (PF) e nos nomes próprios (NOM) não pesada e pesada (FR).

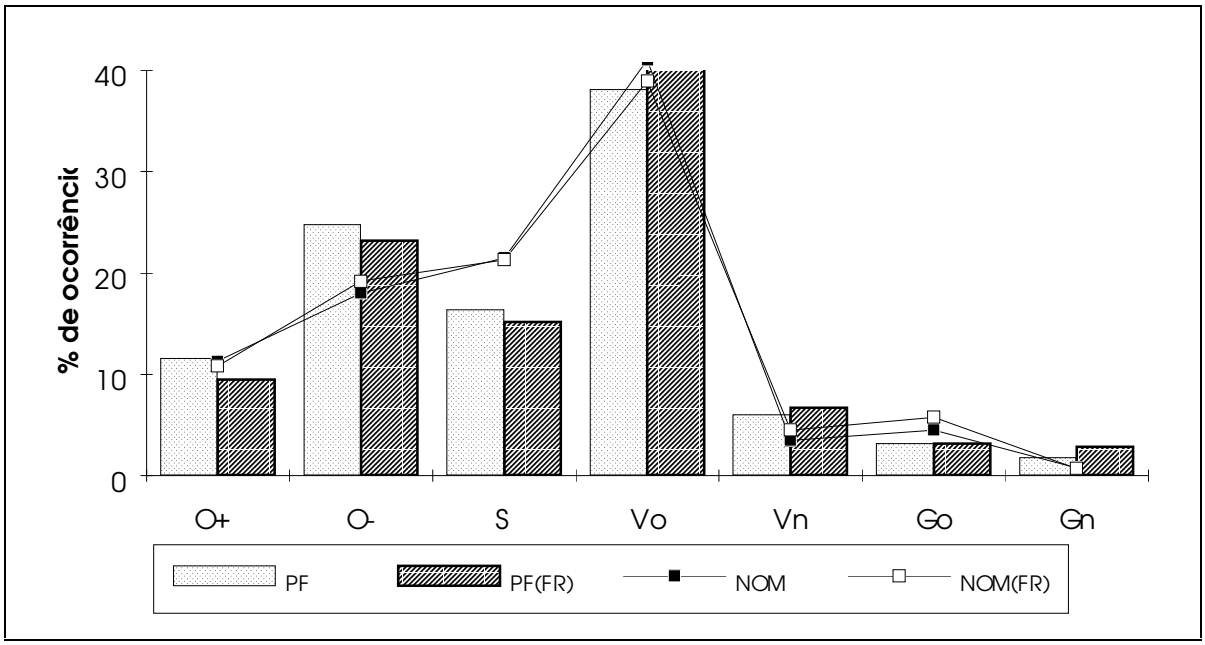

Figura 1 - Distribuição de fones por classes em PF_Fone e Nomes_Fone3 (O=obstruintes; S=consoantes sonantes; V=vogal; G=glide; o=oral; n=nasal; +/-vozeado).

Como a figura 1 mostra, a diferença mais evidente está na maior percentagem relativa de líquidas (grafema "l" e fones [l] e [ɫ]nomeadamente) em nomes próprios. Em termos relativos, contudo, a distribuição das classes de fones é ligeiramente diferente. Enquanto no Português Fundamental a distribuição (pesada e não pesada) é, por ordem decrescente, Oclusiva > Fricativa > Líquida > Nasal > Glide não-nasal > Glide nasal, nos nomes próprios não pesados as líquidas são mais frequentes do que as fricativas. Tendo em consideração apenas as distribuições pesadas, as líquidas são praticamente idênticas às oclusivas e as fricativas ligeiramente superiores a estas. Há, por conseguinte, uma ligeira tendência para uma maior sonoridade dos nomes próprios que se manifesta também por um decréscimo global de 5 a 6% das consoantes [-voz.], como a figura 1 mostra. Proporcionalmente, também, os ditongos orais são mais frequentes nos nomes próprios do que no léxico comum, sendo o inverso verdadeiro para os ditongos nasais.

Em termos de digrafemas e difones, as diferenças são já mais significativas, reflectindo-se na ordenação dos mais frequentes, principalmente quando se trata de análises pesadas que põem em evidência o peso das palavras gramaticais no Português Fundamental e o de primeiros nomes, como "Maria" e “António”, ou apelidos como “Ferreira” e “Pereira”, no corpus de nomes próprios. O mesmo se passa com as distribuições comparadas de trígrafemas e trifones, como o quadro 5 ilustra.

PF		PF (FR)		NOM		NOM (FR)	
DIGRAFEMAS							
ra	1,6	as	1,3	ra	1,7	ar	2,1
ar	1,5	ue	1,2	ar	1,7	ma	1,8
os	1,5	ra	1,2	al	1,5	es	1,8
re	1,5	es	1,2	ca	1,4	ei	1,7
DIFONES							
uʃ	1,5	ɐ̃w̃	1,3	ɐj	1,3	ɐj	1,9
ɐʃ	1,2	ɐʃ	1,1	kɐ	1,3	mɐ	1,8
.iʃ	1,2	d,i	1,0	ɐʃ	1,2	iʃ	1,6
ad	1,1	k,i	1,0	ɾɐ	1,1	ɾɐ	1,6
du	1,1	uʃ	1,0	ɐ ɾ	1,1	ɐ ɾ	1,5
TRIGRAFEMAS							
ent	0,8	que	1,5	eir	1,0	eir	1,4
nte	0,7	não	0,8	inh	0,8	mar	1,1
nte	0,5	ent	0,7	iro	0,6	ira	1,1
ada	0,5	nte	0,5	ira	0,6	ant	0,9
inh	0,5	por	0,5	nha	0,5	ria	0,8
TRIFONES							
adu	0,6	nɐ̃w̃	0,8	ɐjɾ	1,1	ɐjɾ	1,5
adɐ	0,5	paɾ	0,5	jru	0,5	jɾɐ	1,1
muʃ	0,5	pɐɾ	0,4	jɾɐ	0,5	mɐɾ	1,0
mɐ̃t	0,4	umɐ	0,4	jɲɐ	0,4	ɐri	0,9
sɐ̃w̃	0,3	ɐ̃t	0,4	inɐ	0,4	riɐ	0,8

Quadro 5 - Sequências mais frequentes de grafemas e fones e respectivas ocorrências (%) no léxico comum (PF) e nos nomes próprios (NOM) Análise não pesada e pesada (FR), excluindo fronteiras de palavra.

A comparação do número de difones diferentes existentes nos dois corpora resulta num valor superior para o léxico comum: 813 difones face a 791 para os nomes próprios. Há 96 difones do léxico comum que não existem no outro corpus (ex. [dk], [dv], [bs], e alguns difones típicos de formas verbais) e 49 difones do corpus de nomes próprios que não ocorrem em PF_Fone Relativamente aos trifones foram encontrados 8237 para o corpus de léxico comum e 7606 para o corpus de nomes próprios.

Foram também analisadas as distribuições de padrões silábicos e de padrões de palavras para os dois corpora. Os resultados que dizem respeito aos correspondentes a um agrupamento com base em três grandes classes (Consoante, Vogal e Glide), são apresentados nos quadros 6 e 7, respectivamente. Uma parte importante do trabalho realizado no âmbito deste projecto diz respeito aos critérios de silabificação, tendo sido consideradas diferentes análises alternativas e

realizados alguns testes para fundamentá-las. O problema da silabificação é complexo, como é do conhecimento geral, e não é certamente este o momento de discuti-lo em detalhe. Algumas observações a este respeito parecem, no entanto, necessárias para uma melhor compreensão das distribuições aqui referidas e para obviar a algumas dúvidas que estas poderão suscitar quando comparadas com outras, apresentadas anteriormente para esse mesmo corpus [4,20,24].

Na silabificação apresentada nos quadros 6 e 7 assume-se que o português é uma língua em que todos os constituintes da sílaba (ataque e rima) podem ramificar, admitindo no máximo três elementos na rima e sendo o terceiro obrigatoriamente /s/. Na sua generalidade, os critérios seguidos são concordantes com as principais observações em [5,12,13] sobre as sequências de obstruintes, cuja justeza foi apontada em [3,24]. Não são admitidas, no entanto, obstruintes na coda, mesmo nos casos em que se poderiam considerar legitimadas pela presença de um /s/. Formas como *obstáculo* ou *feldspato* são assim sempre transcritas como [ɔ.bʃ'ta.ku.lu] [fɛɬ.dʃ'pa.tu]. Idêntico tratamento foi também o adoptado para casos como *objecto rapto ou hipnose,* cujas transcrições são [ɔ.b'ʒɛ.tu], ['ʁa.p.tu] e [i.p'nɔ.zɨ], respectivamente. Dado que a análise deste tipo de sequências é ainda controversa, considerou-se preferível distingui-las dos verdadeiros grupos consonânticos cujos elementos se associam obrigatoriamente a um mesmo ataque. Pela mesma razão, embora se possa considerar que, de um ponto de vista fonológico, não há ditongos crescentes em português [4,6], estes também são contemplados. Não são aparentes aqui, os casos de ambissilabicidade, discutidos em [4], assumindo-se que as posições de ataque não ocupadas podem sempre ser preenchidas por qualquer elemento que reúna as condições para tal.

No seu conjunto, estes critérios de silabificação permitem descrever alguns aspectos da variação observada quer inter quer intra-locutor e constituem uma base de explicação para algumas das diferenças entre as variantes portuguesa e brasileira (p. ex. ['ʁa.p.tu] (PE) e ['ʁa.pi.tu] (PB)). A grande vantagem destes critérios foi, no entanto, a de permitirem corrigir um número significativo de erros de classificação de nomes próprios como portugueses (ou pronunciáveis como tal) ou como estrangeiros, obedecendo a diferentes princípios de acentuação e de pronúncia.

No quadro 6, em que são apresentadas as percentagens de ocorrência dos diferentes tipos silábicos nos dois corpora, podem assim observar-se alguns padrões inabituais: *C* e *CC*, (como em *rapto* e *adstringente*, respectivamente); $C^{w}V(G)(C)$ em que a consoante associada ao ataque é um /k^{w}/ ou /g^{w}/ (como em *frequência* e *guarda*, por exemplo). Distinguem-se também as sílabas a cujo núcleo estão associadas vogais altas (*V**) e em que o ataque da sílaba seguinte não se encontra preenchido. Certos tipos de sílabas presentes no léxico comum não

ocorrem nos nomes próprios e a sua frequência relativa, não é exactamente idêntica. Algumas dessas diferenças, no entanto, não são significativas e, em qualquer dos casos, mais de 95% das ocorrências dizem respeito aos seis padrões silábicos mais frequentes: CV(e CV*), CVC, V, CVG, CVC e VC, com uma nítida predominância do padrão CV sobre todos os outros. Observam-se, no entanto, algumas discrepâncias no que diz respeito à ordenação relativa destes seis padrões. Estas devem-se, em grande parte, ao peso de palavras gramaticais, como os determinantes *o(s)* e *a(s)* e a partícula de negação *não*, no Português Fundamental, quando pesado. Globalmente, no entanto, os '*ditongos crescentes'* são , mais frequentes nos nomes do que no léxico comum, havendo uma diferença significativa na ocorrência dos padrões CV* nas distribuições pesadas.

	PF	PF(FR)	NOM	NOM(FR)
CCVGC	0,02	<0,01	0,03	<0,01
C^{w}VGC	0,03	0,01	0,00	0,00
V*	0,03	0,01	0,02	<0,01
CC	0,04	0,01	0,00	0,00
CVCC	0,02	<0,01	0,03	0,02
C^{w}VC *	0,09	0,14	0,06	0,02
C^{w}V*	0,16	0,35	0,09	0,02
VGC	0,09	0,07	0,09	0,03
C	0,35	0,11	0,14	0,03
CCV*	0,34	0,19	0,13	0,17
CCVG	0,20	0,10	0,28	0,18
CCVC	0,50	0,39	0,39	0,32
CVGC	0,71	1,31	0,39	0,60
VG	0,53	1,52	0,56	1,05
CCV	4,46	2,82	3,08	2,32
VC	2,76	3,34	2,57	4,31
CVG	5,56	7,41	5,49	5,86
CV*	3,17	1,41	4,11	6,34
V	10,01	17,49	8,15	10,93
CVC	14,65	12,27	11,80	16,18
CV	56,30	51,24	62,58	51,62

Quadro 6 - Ocorrências de padrões silábicos (%) no léxico comum (PF) e nos nomes próprios (NOM), não pesadas e pesadas (FR) (V*=vogal alta seguida de vogal; C^{w}= consoante labializada).

As diferenças são bem maiores quando são considerados padrões de palavras. A variedade dos padrões, como é natural, é muito maior no léxico comum do que nos nomes próprios (2084 e 699 padrões diferentes, respectivamente) e, para um mesmo corpus, o facto de as distribuições serem pesadas, ou não, é determinante.

PADRÃO	PF		PF(FR)		NOM		NOM (FR)	
CV	0,38	(47)	14,94	(1)	0,57	(31)	1,56	(12)
CVCGVC	0,24	(74)	0,16	(56)	0,24	(60)	3,50	(8)
CV$CV	4,51	(2)	8,77	(3)	8,19	(2)	9,46	(1)
CVCVCV	5,69	(1)	2,33	(10)	15,12	(1)	4,98	(5)
CVCVCV$CV	3,09	(5)	0,63	(21)	4,47	(4)	0,83	(25)
CVCVCV$CVC	2,20	(6)	0,30	(35)	0,54	(33)	0,06	(122)
CVCVCVC	4,26	(3)	1,19	(16)	3,81	(5)	1,10	(17)
CVCVCVG	1,98	(7)	0,44	(28)	0,80	(22)	0,10	(96)
CVCVV	0,30	(60)	0,38	(30)	0,36	(51)	4,61	(6)
CV$CVC	3,51	(4)	4,10	(7)	4,69	(3)	8,60	(2)
CV$CVG	1,71	(9)	1,36	(15)	1,56	(9)	0,33	(54)
CVCVGCV	0,35	(50)	0,35	(31)	2,09	(7)	5,64	(3)
CVC	0,42	(41)	4,71	(6)	0,43	(43)	0,43	(44)
CVC$CV	1,32	(12)	2,35	(9)	2,06	(8)	5,40	(4)
CVCCVCV	1,54	(11)	0,89	(18)	3,33	(6)	3,48	(9)
CVC$CVC	1,03	16	0,55	(24)	1,22	(17)	3,42	(10)
CVG	0,27	(66)	6,32	(4)	0,44	(41)	0,37	(51)
V	0,10	(140)	14,78	(2)	0,05	(183)	0,72	(35)
V$CV	0,47	(39)	5,44	(5)	0,55	(32)	0,42	(45)
VCVCG$V	0,12	(132)	0,03	(151)	0,83	(20)	3,81	(7)
VCVCV	1,68	(10)	1,41	(14)	1,50	(10)	1,25	(14)
VCVCV$CV	1,80	(8)	0,32	(34)	1,38	(12)	0,76	(31)
VG	0,05	(223)	2,37	(8)	0,03	(215)	< 0,01	(484)

Quadro 7 - Os 10 padrões de palavra mais frequentes no lexico comum (PF) e nos nomes próprios (NOM) não pesados e pesados (FR). Entre parêntesis, indica-se o número de ordem no corpus (padrão mais frequente = 1).

Dado o grande número de padrões possíveis, apenas se apresentam no quadro 7 os dez mais frequentes para cada um dos corpora. O número de ordem do padrão nas diferentes situações é também indicado para dar uma ideia das principais diferenças entre as formas do léxico comum e dos nomes próprios e ainda para mostrar a variação no interior de cada corpus em função da frequência. É digno de nota que apenas 3 dos padrões (CV$CV, CV$CVC e CVCVCV) se encontrem entre os 10 mais frequentes em qualquer circunstância. A presença de monossílabos entre os 10 padrões mais frequentes reflecte, evidentemente, o peso das palavras gramaticais no Português Fundamental. A relação entre a frequência de ocorrência de uma forma e a sua extensão em número de sílabas não é, contudo, idêntica nos dois corpora, como está melhor ilustrado na figura 2(a).

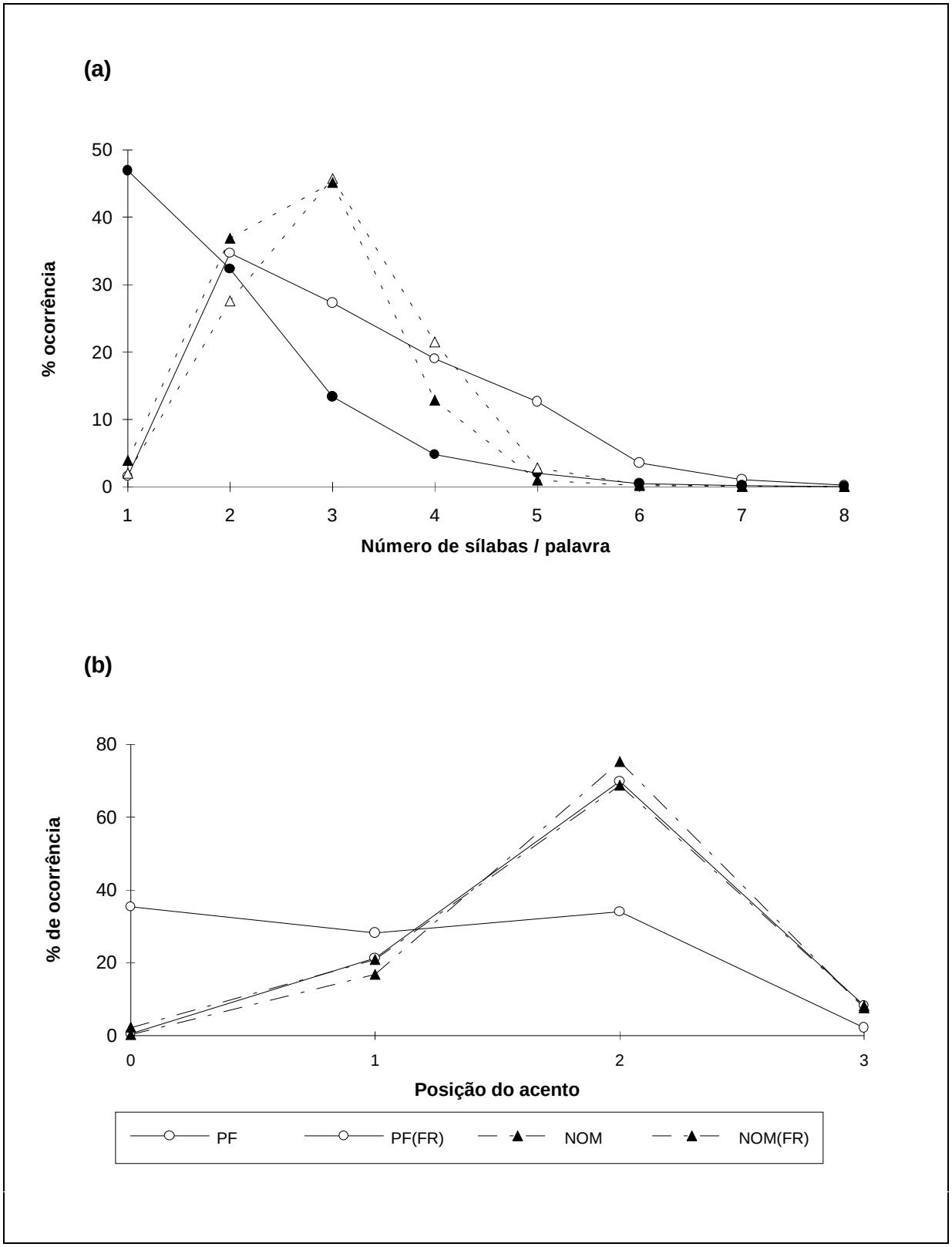

Figura 2 - Distribuição das formas em função de (a) número de sílabas e (b) posição do acento no lexico comum (PF) e nos nomes próprios (NOM) não pesada e pesada (FR). A posição do acento é contada a partir do fim da palavra (0 - não acentuada; 1 - oxítona; 2 - paroxítona; 3 - preparoxítona).

nsil	Sil_ac	PF	PF(FR)	NOM	NOM(FR)
9	1	<0,01	<0,01	0,00	0,00
10	2	<0,01	<0,01	0,00	0,00
8	1	0,01	<0,01	0,00	0,00
9	2	0,02	<0,01	0,00	0,00
9	3	0,02	<0,01	0,00	0,00
3	0	0,03	0,03	0,00,	0,00,
8	3	0,06	<0,01	0,00	0,00
8	2	0,15	0,01	0,02	<0,01
7	3	0,29	0,04	0,02	0,01
7	2	0,68	0,11	0,03	0,01
2	0	0,20	6,98	0,00,	0,00
6	3	0,90	0,10	0,13	0,02
6	1	0,28	0,03	0,05	0,07
7	1	0,06	<0,01	0,01	0,10
6	2	2,32	0,31	0,16	0,10
5	1	1,40	0,20	0,12	0,15
5	3	2,21	0,42	1,03	0,26
4	1	4,15	0,75	0,80	0,48
5	2	9,02	1,37	1,68	0,58
1	1	1,30	18,48	1,90	1,65
3	3	1,82	0,87	2,18	2,00
1	0	0,27	28,43	0,12	2,23
4	3	3,05	0,80	4,26	5,84
4	2	20,08	3,25	16,45	6,52
3	1	7,99	1,99	5,10	7,67
2	1	6,06	6,81	8,85	10,80
2	2	12,76	18,54	18,56	26,02
3	2	24,86	10,47	38,43	35,47

Quadro 8 Distribuição das formas do corpus em função do número de sílabas (nsil) e da posição do acento (Sil_ac) no léxico comum (PF) e nos nomes próprios (NOM) não-pesada e pesada em frequência (FR). A posição do acento é contada a partir do fim da palavra (0 - não acentuada; 1 - oxítona; 2 - paroxítona; 3 - proparoxítona).

Considerando apenas as distribuições pesadas, verifica-se que a frequência de ocorrência dos dissílabos é muito semelhante mas, enquanto no Português Fundamental a frequência diminui claramente à medida que o número de sílabas aumenta, já não é assim para os nomes próprios, em que há uma preferência clara pelos trissílabos. É de notar, ainda, que enquanto cerca de 95% dos nomes próprios tem entre 2 a 4 sílabas, as formas do léxico comum com a mesma extensão apenas correspondem a 50,5% de PF_Fone.

Como é bem conhecido, a grande maioria das palavras do português são acentuadas na penúltima sílaba. Esta tendência pode observar-se na figura 2(b) tanto para o léxico comum como para os nomes próprios mas, também aqui, as

distribuições pesadas mostram algumas diferenças significativas: para os primeiros, a distribuição das palavras gramaticais, inerentemente não acentuadas, das oxítonas e das paroxítonas é bastante mais equitativa do que para os segundos (35,4%, 28,3% e 34,1%, contra 2,3%, 20,9% e 68,7%, respectivamente). Os nomes mais frequentes em português são, então, paroxítonos e trissilábicos. O quadro 8 mostra as distribuições observadas quando o número de sílabas e a posição do acento são consideradas simultaneamente.

2. Transcrição fonética automática

Foram duas as metodologias testadas para a transcrição fonética automática dos diferentes corpora. A primeira, desenvolvida no âmbito do projecto DIXI, consiste num sistema de regras. Todo o código foi programado em linguagem C, directamente no caso da atribuição do acento, e com base no compilador SCYLA [8], para as restantes regras. A estrutura multi-dimensional deste compilador permite a cada procedimento ter acesso simultâneo a todos os resultados dos procedimentos anteriores. Apresenta ainda as vantagens de gerar código C e de ter uma ferramenta poderosa para teste e correcção das regras. O sistema permite diferentes estilos de transcrição e pode colocar as marcas de acentuação quer antes do ataque quer antes do núcleo da sílaba. São 18 as regras de atribuição do acento utilizadas. A taxa de erros resultante é muito baixa, devendo-se, na maior parte dos casos à supressão de marcas gráficas indicadoras de um acento secundário na base quando, em formas derivadas por sufixação, esta não é desacentuada e se torna impossível desfazer a ambiguidade resultante dessa supressão sem recorrer a um dicionário de excepções. Segue-se, já implementado sobre o SCYLA, um módulo de silabificação e um módulo de transcrição fonética com cerca de 200 regras.

O segundo tipo de método baseia-se numa rede neuronal. São várias as propriedades que caracterizam as redes neuronais e que justificam a sua designação por analogia com o sistema nervoso: capacidade de aprendizagem, extracção de características, generalização e processamento paralelo. Sendo estas propriedades obviamente importantes no processo da leitura, não é de estranhar que os primeiros trabalhos de aplicação de redes à conversão grafema-fone datem já de 1987, altura em que Sejnowski apresentou pela primeira vez o sistema conhecido por NETTALK [17]. Tal como neste trabalho precursor, a rede adoptada é do tipo multi-camada, treinada pelo algoritmo de retropropagação de erro.

O treino da rede é feito através de uma aprendizagem supervisionada em que, à entrada da rede, é apresentado o grafema a transcrever, rodeado pelo seu contexto,

sendo especificada qual a saída pretendida. A rede "aprende" ajustando os pesos das ligações entre as várias unidades de processamento ou neurónios.

O processo de treino, no entanto, deve ser precedido por uma etapa de alinhamento grafema-fone do corpus, uma vez que a cada símbolo de entrada (grafema) nem sempre corresponde apenas um símbolo de saída (fone) e vice-versa. Torna-se necessário indicar que certos grafemas não têm realização fonética (caso do “h” inicial, por exemplo), que a uma sequência de grafemas pode corresponder um só fone, (ex. dígrafos) e que a um só grafema pode corresponder uma sequência de fones (ex. ditongos que correspondem a grafemas simples). O alinhamento foi efectuado automaticamente, através da adaptação de programas desenvolvidos no âmbito do projecto DIXI.

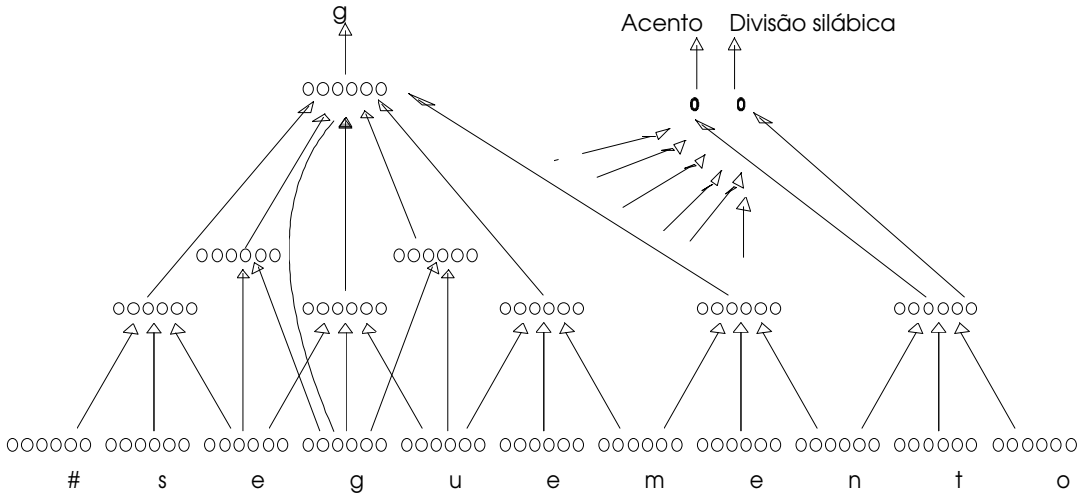

Figura 3 - Arquitectura da rede multicamada

A arquitectura adoptada para a rede está esquematizada na Figura.3. A camada de entrada é constituída por 11 grafemas: o grafema a transcrever, 3 grafemas à esquerda e 7 grafemas à direita, dos quais apenas 5 são utilizados para a transcrição fonética, sendo os restantes apenas necessários para efeitos de acentuação. A cada grafema correspondem 36 entradas, uma por cada um dos 36 grafemas diferentes (contando separadamente os grafemas com diacríticos e o símbolo gráfico de fronteira de palavra), o que perfaz um total de 396 entradas.

A camada escondida está estruturada em 5 grupos de trigrafemas e 2 grupos de digrafemas (incluindo os grafemas imediatamente à esquerda e à direita do grafema a transcrever), sendo cada grupo constituído por 20 unidades, num total de 140 unidades escondidas.

Existem 47 unidades de saída, uma por cada uma das 45 unidades fonéticas consideradas (incluindo unidades simples e complexas), uma para a marca do acento de palavra (acento principal) e outra para a marca de fronteira de sílaba.

Nesta fase, o acento secundário não foi contemplado com uma marca específica, só podendo ser acedido indirectamente nas formas em que se reflecte na qualidade vocálica. De modo a diminuir o número de pesos a ajustar, adoptaram-se pesos partilhados (11087 pesos para as 21167 sinapses). Existe ainda uma ligação directa da entrada para a saída.

A rede foi treinada com um subconjunto aleatoriamente seleccionado do corpus PF_Fone (cerca de 70%, num total de 100.000 fones). Ao fim de 8 iterações, o erro ao nível do segmento era já de 1,5%, baixando para 1% ao cabo de 40 iterações. Ao fim de um dia de treino, os resultados já eram significativamente parecidos com os indicados nas terceiras colunas do Quadro 9 a) e b) para os corpora PF_Fone de teste (os restantes 30%) e Nomes_Fone2, respectivamente.

A percentagem de erros de transcrição fonética ao nível da palavra efectuados pelo sistema de regras está indicada nas segundas colunas dos Quadros 9 (a) e (b), respectivamente para o subconjunto PF_Fone de teste e para o corpus Nomes_Fone2. Indicam-se também as percentagens de erros de atribuição do acento principal e de silabificação.

	TIPO DE ERRO	REGRAS	REDE
(a)	Transcrição fonética	4,5 %	7,3 %
	Acentuação	0,4 %	2,7 %
	Silabificação	0,3 %	0,8 %
(b)	Transcrição fonética	7,3 %	12,4 %
	Acentuação	0,4 %	1,1 %
	Silabificação	0,3 %	1,0 %

Quadro 9- Comparação do desempenho do sistema de regras e da rede neuronal: (a) Corpus PF_Fone de teste; (b) Nomes_Fone2.

Da observação destes valores pode concluir-se que, para o Português, ao contrário do que é muitas vezes referido para outras línguas, o desempenho das regras na transcrição de nomes próprios é apenas ligeiramente inferior ao observado para o léxico comum, sendo o das redes inferior ao das regras em cerca de 3% e 5%, respectivamente. Note-se, no entanto que, no que diz respeito à atribuição do acento, os valores apresentados no quadro 8 para a rede neuronal não correspondem a uma análise das transcrições na saída. De facto, a rede tende a atribuir uma multiplicidade de acentos a uma mesma palavra e foi necessário um pós-processamento para reter apenas o mais à direita.

Para o primeiro corpus, verificou-se que 59% das palavras em que o sistema de regras falha são também erradamente transcritas pela rede neuronal e que os erros

cometidos por ambos os métodos são rigorosamente idênticos em 44% dos casos. Para o segundo corpus, estes valores são ainda mais elevados: 74% e 56% respectivamente.

A maior parte dos erros comuns tem lugar na transcrição dos grafemas *e* e *o* que, como já foi apontado anteriormente, são os que levantam um maior número de problemas (cf. [20]). A rede, no entanto, parece ter maior dificuldade em lidar com a nasalidade: uma consoante nasal em posição final de sílaba nem sempre nasaliza a vogal precedente e uma lateral nessa mesma posição fá-lo por vezes. A nasalização pode também ocorrer em certos casos em que a rede associa a consoante ao ataque da sílaba seguinte.

Enquanto as regras se equivocam, sistematicamente, em casos como *padeiro*, em que ocorrem vogais átonas não elevadas, desde que estes não estejam incluídos na lista de excepções, a rede nem sempre o faz. Em contrapartida, prediz estranhas elevações de vogais acentuadas e nem sempre eleva as vogais átonas. Parece ter, também, alguma dificuldade com a análise de sequências vocálicas, interpretando como hiatos alguns ditongos e vice-versa.

Parte dos erros cometidos pelas redes podem explicar-se pela insuficiente representatividade de certas sequências de grafemas no corpus de treino. É de reparar, no entanto, que o facto de a rede atribuir múltiplos acentos a uma mesma palavra pode também contribuir para uma maior inconsistência na interpretação de sequências de vogais e na predição da qualidade destas.

Convém referir ainda que o desempenho do sistema de regras para os nomes próprios poderá aproximar-se do observado para o léxico comum se forem introduzidas pequenas modificações que permitam lidar com as consoantes geminadas das grafias conservadoras e com a presença de obstruintes em posição final absoluta, evitando, por exemplo, que a nomes como *David* seja atribuído o acento à penúltima vogal. Estas modificações são evidentemente necessárias para melhorar as transcrições automáticas das siglas e dos nomes de origem estrangeira.

3. Nomes de empresas e serviços públicos

Os resultados acima mencionados não contemplam os nomes de empresas e de serviços públicos que constituem cerca de 33% das entradas da base de dados dos TLP e para os quais tanto o desempenho das regras como o da rede neuronal se revelaram claramente insatisfatórios (apenas 57% e 49%, de resultados coincidentes com as transcrições fonéticas manuais, respectivamente).

Estes nomes apresentam, de facto, um conjunto de particularidades que tornam a sua leitura difícil, mesmo por parte dos falantes nativos, como se verificou imediatamente pelas dúvidas e oscilações de pronúncia que surgiram durante a

fase de correcção manual. A inconsistência das transcrições manuais constitui, naturalmente, uma dificuldade adicional que impede, por um lado, uma medida objectiva do desempenho dos métodos automáticos e dificulta, por outro, a escolha das estratégias a adoptar para o seu processamento automático, tanto no que diz respeito às regras de conversão grafema-fone como à constituição de um corpus para treino da rede. Foi necessário, por conseguinte, realizar um estudo mais aprofundado, em que se procurou fazer um levantamento dos principais problemas e encontrar explicações para a variação observada.

3.1. Comparação das transcrições automáticas e manuais

Tal como para as outras categorias de nomes, o primeiro passo consistiu na comparação das transcrições fornecidas pelos dois métodos automáticos com as sugeridas pelo(s) transcritor(es) durante a fase de correcção manual. Para além dos erros já referidos para as outras categorias, e que são naturalmente persistentes, foram observadas múltiplas discrepâncias entre as transcrições automáticas e as manuais.

Parte dessas discrepâncias devem-se ao facto de nem sempre serem respeitados alguns preceitos ortográficos básicos tanto no que diz respeito tanto à correspondência grafema-fone como à colocação das marcas gráficas de acento. Uma forma como *alfasom*, por exemplo, é naturalmente transcrita como [aɫ.fɐ'zõ] porque, em português, um *s* em posição intervocálica se deveria pronunciar invariavelmente como [z]. A frequente omissão de diacríticos indicadores da posição do acento, como em *tecnindustria* (em vez de *tecnindústria*), conduz, por sua vez, a transcrições como [tɛ.k.nĩ.duʃ 'tri.ɐ] Um número considerável de formas como as terminadas em -*ax*, *-ux* ou *-trans*", que são sempre oxítonas, são interpretadas como paroxítonas pelos dois métodos, uma vez que sendo raras ou inexistentes no léxico comum não foram aprendidas pelas redes nem contempladas pelo módulo de acentuação de DIXI que, por defeito, lhes aplica a lei geral.

A maior parte das discrepâncias dizem respeito, no entanto, à dificuldade (vd. incapacidade) em predizer adequadamente as realizações fonéticas das vogais átonas para as formas contidas neste corpus e devem-se não só à atribuição do acento mas também ao facto de nem sempre a elevação das vogais átonas ter lugar. Vejam-se, apenas a título de exemplo, as diferentes transcrições obtidas para uma forma como *jovali*: [ʒu.vɐ 'li] (regras); [ʒu.'va.li] (rede) e [ʒɔ'va.li] ou [ʒɔ.va'li] (correcções manuais alternativas). O sistema de regras produz uma transcrição que se pode considerar perfeitamente correcta, em função das regras gerais da ortografia e da pronúncia, à semelhança, aliás, do que faz para *javali* que é uma palavra comum e se transcreve como [ʒɐ.vɐ 'li]. Diferenças do mesmo género entre

as transcrições automáticas e manuais mostram, também, que a elevação das vogais pode não ter lugar, mesmo quando o ataque da sílaba seguinte é preenchido por uma consoante nasal, como em *granitex*, forma automaticamente transcrita como [grɐ'ni.tɛks] e manualmente como [grɐ.ni'tɛks] ou [gra.ni'tɛks]. Contrariamente ao que se observa para as outras categorias de nomes, há uma grande variação nas pronúncias sugeridas ou aceites pelos diferentes transcritores. Um dos exemplos paradigmáticos é *gravatex* forma para a qual foram consideradas aceitáveis todas as combinações de realizações fonéticas possíveis das duas vogais à esquerda do acento: [grɐ.vɐ'tɛks], [grɐ.va'tɛks], [gra.vɐ'tɛks] e [gra.va'tɛks].

Estas oscilações de pronúncia podem encontrar uma explicação no facto de serem possíveis diferentes análises para esta sequência segmental: a terminação, que pode ser considerada típica desta classe de nomes, tanto pode ser *-ex* como -t*ex* (truncamento de *texto* ou de *têxtil*) e o primeiro elemento tanto pode ser uma palavra (*grava*) como um radical (*gravat*). Globalmente, a forma pode ser interpretada como "derivada" ou como "composta", recebendo um ou dois acentos, respectivamente. Não sendo (pela sua terminação) uma palavra do português, existe ainda uma outra leitura alternativa em que as vogais átonas à esquerda do acento principal não sofrem qualquer elevação. A não-elevação das vogais átonas à esquerda do acento pode ser, no entanto, apenas encarada como uma tendência ou uma alternativa possível para um número significativo de formas, mas não para todo o tipo de formas presentes neste sub-corpus, como ilustram as pronúncias de *alfasom* e *copicanola*: [aɫ.fɐ'sõ] e [kɔ.pi.kɐ'nɔ.lɐ̃].

Exemplos deste tipo sugerem a existência de alguma relação entre a análise que é feita da estrutura interna das palavras e a sua pronúncia, mostrando, também, que essa análise está sujeita a oscilações cuja origem e fundamento nem sempre são evidentes. Procurou-se, assim, fazer uma pesquisa tão exaustiva quanto possível do tipo de elementos e de processos que intervêm na constituição dos nomes em Acro_Fone.

3.2. Elementos e processos lexicais utilizados

Para a pesquisa de elementos constituintes, todas as entradas foram novamente comparadas não só com o conjunto de formas de citação e formas flexionadas do dicionário já utilizado para a sua classificação inicial, mas ainda com as do subcorpus Nomes_Fone3. O objectivo foi o de verificar, para cada entrada, se existiam sequências segmentais identificáveis como formas do léxico comum, nomes, apelidos ou topónimos. Todas as entradas de Acro_phone foram ainda comparadas entre si para pesquisa de sequências recorrentes de 3 ou mais letras, cuja frequência de ocorrência foi também calculada. Chegou-se, assim, a um conjunto de potenciais elementos constituintes que foi, em seguida, verificado manualmente. Nessa verificação foi naturalmente tido em conta o grau de

ambiguidade intrínseca de cada sequência. Por exemplo, uma sequência como -*trónica* é considerada não ambígua uma vez que não corresponde ao truncamento de qualquer nome de baptismo, apelido ou topónimo e que apenas ocorre nas palavras *electrónica* e *neutrónica*. Para outras sequências, no entanto, o grau de ambiguidade pode ser demasiado elevado para justificar que sejam retidas como constituintes.

MODIFICADORES MORFOLÓGICOS	HIPER	*HIPER*MERCADO
	INTER	*INTER*LAR
	POLI	*POLI*GRUPO
PALAVRAS COMUNS	CONSUL(TA)	DIGI*CONSUL*
	GESTÃO	AGRO*GEST*
	SISTEMAS	*SIS*NORTE
NOMES/ APELIDOS	GOMES	TECNO*GOMES*
	LOPES	PUBLI*LOPES*
	ABÍLIO	*ABILI*MÓVEIS
TOPÓNIMOS	ALGAR(VE)	*ALGAR*OTEL
	GONDO(MAR)	*GONDO*PREDIAL
	LIS(BOA)	*LIS*FRIO
EMPRÉSTIMOS	PRESS	UNI*PRESS*
	TRADE	PLANI*TRADE*
	TOUR	CABI*TOUR*

QUADRO 10 - Tipos de elementos constituintes mais frequentes em nomes de empresas.

Considerando apenas as sequências com frequência igual ou superior a 10 e não eliminadas manualmente devido à sua ambiguidade, foi possível constituir uma lista de 660 elementos, de que o quadro 10 apresenta alguns dos exemplos mais frequentes. Apesar das suas dimensões, este conjunto de elementos assegura uma cobertura razoável do corpus Acro_Fone: 15% das entradas são totalmente cobertas pela justaposição de elementos (2% com haplologia); 50% das entradas são parcialmente cobertas (23% apresentam o elemento na posição inicial, 20% na posição final e 7% em ambas as posições); os restantes 35% são siglas ou acrónimos que não são cobertos por este conjunto.

Na constituição dos nomes de companhias, são utilizados, modificadores morfológicos[3] radicais e palavras do léxico comum, primeiros nomes, apelidos, topónimos e praticamente todas as possíveis abreviaturas destes. Todos estes elementos podem ser livremente combinados entre si, com palavras estrangeiras ou com terminações características desta classe de nomes.

Em Acro_Fone, existem, assim, numerosos exemplos de nomes de companhias com origem em processos de criação lexical utilizados em português [9], tais como a

[3] Adoptou-se aqui a designação proposta em [23] para constituintes cuja categoria morfológica é complicada de determinar.

acronímia, a amálgama, a sigla e também, embora muito raramente, o truncamento. Trata-se, na maior parte dos casos, de abreviaturas da designação geral da empresa ou de um ou mais nomes e/ou apelidos do(s) seu(s) proprietário(s). Essas abreviaturas podem incluir apenas a letra inicial de cada uma (sigla), uma ou mais letras, sílabas ou mesmo morfemas iniciais (acrónimo) ou qualquer sequência de elementos aleatoriamente seleccionados (amálgama) [9].

A distinção entre estes diferentes processos de criação lexical nem sempre é clara. Uma forma como *anarec*, por exemplo, pode ser uma abreviatura de *Ana Rebelo Castro*, de Abel *Neves Alves: Restauro de Embutidos e Carpintaria*, de *Associação Nacional dos Amigos das Reservas Ecológicas Costeiras* ou de qualquer outra combinação de uma ou mais letras iniciais de nomes próprios ou de palavras do léxico comum. Trata-se, de facto, *da Associação Nacional de Revendedores de Combustíveis* que é um acrónimo e não uma sigla, uma vez que nem sempre foi retida apenas a primeira letra de cada palavra. Como tem sido apontado em alguns dos trabalhos que se ocupam desta classe de nomes [15,16,25], a distinção fundamental não está propriamente no número de letras que é retido mas nos critérios que estão na base da sua selecção: enquanto os acrónimos são sempre construídos para serem "lidos" as siglas podem ser lidas ou soletradas, justificando-se algumas apenas pela facilidade de escrita.

Encontra-se também, naturalmente, um grande número de entradas com origem em processos ditos de formação lexical mas, ao contrário do que se observa para o léxico comum, a composição é um processo extremamente produtivo[4] neste tipo de nomes. Ora, são os compostos graficamente aglutinados e os derivados em *-mente* e *Z-avaliativos* (cujo estatuto é ambíguo) as formas que, devido à sua dupla acentuação, exigem um tratamento especial e justificam as poucas regras de base morfológica contempladas no sistema DIXI (cf. [19]). A análise morfológica necessária para conseguir uma leitura adequada da maior parte dessas formas é muito pouco elaborada e apenas faz apelo a um conjunto reduzido de radicais e de afixos. Este tipo de tratamento só é possível porque a ortografia portuguesa trata os compostos de um forma que pode ser ambígua de um ponto de vista morfo-sintáctico ou semântico (cf. [7]) mas que tem como objectivo central garantir uma leitura adequada[5]. Desse ponto de vista, parece fundamental a distinção feita em

[4] É bem possível que não se trate de uma característica específica deste tipo de nomes e que a produtividade do processo de composição seja bem maior no português actual do que, em geral, se pensa que é. A este respeito, vejam-se [1] e [23].

[5] A preocupação com a facilidade de leitura é constante e bem clara relativamente aos compostos "*Emprega-se o hífen nos compostos em que entram,* ***fonéticamente distintos (e, portanto, com acentos gráficos, se os têm à parte****), dois ou mais substantivos, ligados ou não por preposição ou outro elemento, um substantivo e um adjectivo, um adjectivo e um substantivo, dois adjectivos ou um adjectivo e um substantivo com valor adjectivo, uma forma verbal e um substantivo, duas formas verbais, ou ainda outras combinações de palavras, e em que o conjunto dos elementos, mantida a noção de composição, forma um sentido único ou uma aderência de sentidos [...] Se,*

[22] entre compostos de radicais (CRs) e compostos de palavras (CPs). Em português, não há desacentuação dos radicais e, como a elevação é um fenómeno específico das vogais átonas, podem ocorrer nestas formas pelo menos tantas vogais abertas quantos os elementos constituintes. Nos compostos de radicais (CRs), no entanto, existe uma vogal de ligação (/i/ ou /ɔ/) que, quando /ɔ/, também não sofre qualquer elevação. Estes compostos podem apresentar, por conseguinte, mais uma vogal aberta do que os CPs equivalentes, uma vez que restrições idênticas se não aplicam às marcas de género [22]. A ortografia portuguesa distingue estes dois tipos de compostos, aglutinando os CRs numa só palavra gráfica[6] e tratando os outros como sequências de duas ou mais palavras independentes, separadas entre si por espaços ou por hífenes. Todos os casos de CPs que se escrevem como uma só palavra gráfica, como *pontapé*, *varapau* e *pernalta* correspondem a formas que já se não podem considerar como compostas, podendo apresentar apenas uma única vogal não-elevada: a acentuada do último elemento[7]. Os CPs não necessitam, por conseguinte de qualquer tratamento especial para serem correctamente transcritos e a maior parte dos CRs é identificável com base numa lista relativamente reduzida de morfemas presos, na sua maioria de origem greco-latina. Para as formas em Acro_Fone, um tratamento deste tipo é claramente inadequado, uma vez que, independentemente do seu tipo, todos os compostos são graficamente aglutinados e as marcas gráficas de acento estão frequentemente ausentes. São dificuldades adicionais que contribuem para o desempenho insatisfatório dos sistemas automáticos e que também podem explicar parte das dúvidas e oscilações durante a fase de correcção manual.

3.3. Relação entre os processos lexicais e o comportamento dos falantes

Para estudar a variação na pronúncia destes nomes por parte dos falantes e procurar relacioná-la com os processos lexicais utilizados para os construir, foram recolhidas informações complementares: (I) directamente junto de um conjunto de empresas para averiguar qual a origem e pronúncia dos seus nomes; (2) junto de 10

---

*porém, no conjunto de elementos de um composto está perdida a noção de composição, faz-se a aglutinação completa" " (Base XXVIII do Acordo ortográfico de 1945). "Emprega-se o hífen em palavras formadas com prefixos gregos de origem grega ou latina, ou com outros elementos análogos de origem grega (primitivamente adjectivos), quando convém não os aglutinar aos elementos imediatos, por motivo de clareza ou expressividade gráfica**, por ser preciso evitar má leitura**, ou por tal ou tal prefixo ser acentuado graficamente*.

[6] Repare-se que embora as instruções para a organização do Vocabulário Ortográfico Resumido da Língua Portuguesa, determinem que os elementos de compostos de adjectivos são sempre separados por hífen, mesmo quando são utilizadas as formas reduzidas (como em agro-pecuário, nipo-soviético, etc), estas normas raramente são seguidas: o uso tende a aglutiná-los graficamente, tratando-as como quaisquer outros CRs.

[7] Há um pequeno conjunto de formas que constituem uma excepção a esta regra na medida em que a vogal correspondente à acentuada do primeiro elemento não se eleva ou pode apresentar ainda oscilações de pronúncia (ex.*madrepérola, passaporte, clarabóia, rodapé, regabofe*,etc).

falantes de formação escolar de nível universitário, a quem foi pedida a leitura de uma lista de 100 itens, aleatoriamente seleccionados de entre as entradas do corpus Acro_Phone e não anunciados na comunicação social.

O contacto directo com as empresas mostrou, sobretudo, a grande variedade dos critérios que podem presidir à escolha de um nome. Em termos gerais, pode pretender-se que a forma resultante soe como autóctone ou como estrangeira, que seja homógrafa (ou homófona) de uma palavra do léxico comum ou totalmente distinta destas. Pode ainda pretender-se favorecer ou desfavorecer certas associações semânticas ou, simplesmente, evitar que o nome escolhido seja idêntico ou muito semelhante a outro já existente. As reacções em relação às perguntas directamente relacionadas com a pronúncia de certos nomes foram, muitas vezes, de surpresa: "*pois não é evidente como é que o nome se lê*?" Só que essa evidência pode ser que se *"deve ler como se lê em português*" ou que se "*deve ler como uma sigla*". E '*ler como uma sigla"* tanto pode significar que os elementos constituintes mantêm a pronúncia que tinham nas formas de que foram extraídos como que não se elevam quaisquer vogais à esquerda do acento. Assim, por exemplo, uma sequência inicial *art* com origem no nome *artur* tanto se pode ser lida [ɐrt] como [art], pelos proprietários da empresa, mesmo quando esta nada tem a ver com o ramo artístico. A segunda leitura é, no entanto, a única que se obtem se essa relação existir.

Resolver as dúvidas e as oscilações de pronúncia por inquérito directo junto das empresas não é praticável nem sequer adequado: o que se pretende obter não são apenas as pronúncias que os proprietários imaginaram para as suas empresas, mas as que correspondem à sua leitura mais provável pelos falantes de português. Por outras palavras, o que se pretende simular é o comportamento de um operador humano e a grande questão está em saber quais são as pronúncias possíveis e mais prováveis e quais as que são declaradamente inaceitáveis.

O facto de, no teste de leitura, apenas 37% das produções dos falantes serem concordantes entre si mostra bem a extrema variabilidade de pronúncia a que estas formas estão sujeitas. Uma análise mais cuidada permite mostrar, no entanto, que a variação não é aleatória.

Muitas das formas presentes no corpus são inequivocamente analisadas como compostas (ex. *globomar* e *frangolândia*). Dado que a vogal de ligação dos CRs é graficamente idêntica à marca do masculino ("o") dos CPs e que todos os compostos são aglutinados, as formas deste tipo são inerentemente ambíguas e prestam-se a oscilações de pronúncia. Estas oscilações prendem-se, no entanto, apenas com a dificuldade dos falantes em identificar o tipo de composto que está em causa. Assim, *globomar*, por exemplo, foi pronunciada como [glo.bɔ'mar] e [glo.bu'mar] em 40% e 60% dos casos respectivamente, mas não se observam

realizações do tipo *[glu.bu'mar] ou *[glɔ.bɔ'mar] que são consideradas como inaceitáveis.

As oscilações de pronúncia dos falantes apontam, no entanto, para um tipo de dificuldade mais geral: o reconhecimento de palavras ou de radicais dentro de palavras gráficas não parece ser uma tarefa que faça parte dos hábitos de leitura dos portugueses. Se o fosse, seria de esperar que formas graficamente não ambíguas como *alfasom*, *macara*, e *mataratos* fossem invariavelmente interpretadas como CPs e pronunciadas como [aɫ.fɐ'sõ], [ma'karɐ] e [matɐ'ʀatuʃ], respectivamente. Elas são, no entanto, preferencialmente interpretadas como palavras simples e lidas como [aɫ.fɐ'zõ] , [mɐ'karɐ] e [mɐtɐ'ratuʃ], em 60% dos casos para a primeira destas formas e em 100% para as duas últimas. O comportamento dos falantes parece assim apontar para um processamento do tipo do que foi adoptado para o sistema de regras e que inclui a pesquisa de um elemento inicial, em geral bi ou trissilábico, que termine em /i/ ou /ɔ/. A dificuldade da tarefa de reconhecimento de radicais ou de palavras no interior de palavras gráficas explica, pelo menos em parte, a inconsistência na pronúncia de "s" e "r" em posição intervocálica: [s] e [ʀ] quando estes são analisados como primeiro elemento de um constituinte não inicial de um composto e [z] e [r] sempre que a sequência é interpretada como uma palavra simples, seguindo as regras gerais de pronúncia. Repare-se, contudo, que a análise da estrutura interna de uma palavra gráfica pode ser condicionada, pelo menos em parte, pela ambiguidade da própria grafia. Dois dos informantes, embora tenham posto a hipótese de *alfasom* poder ser um composto de palavra, rapidamente a afastaram, uma vez que, se assim fosse, se deveria escrever *alfassom*, à semelhança do que acontece com outras formas com origem em processos de composição como, por exemplo, *madressilva*.

Os falantes fornecem espontaneamente mais do que uma leitura para certas formas. Para *bitolagrafica*, por exemplo, a primeira leitura foi frequentemente silabada (ex. [bi.tu.lɐ.grɐ'fi.kɐ]). A localização do acento de palavra foi quase sempre corrigida numa segunda leitura ([bi.tu.lɐ'gra.fi.kɐ]), mas a leitura desta forma como composta ([bi.tɔ.lɐ'gra.fi.kɐ])apenas ocorreu como terceira leitura e apenas em 50% dos casos., apesar de todos os informantes conhecerem a palavra *bitola*. De um modo geral, no entanto, têm consciência de que os nomes de empresas e de serviços públicos diferem das formas do léxico comum e dos nomes próprios, tanto na grafia como na pronúncia. Assim, à medida que se apercebem qual é a classe de nomes que está em jogo, passam a querer analisar, sempre que possível, todas as formas como compostas, atribuindo um acento a cada elemento que coincida com um radical ou com uma palavra ou que possa ser interpretado como um truncamento de qualquer deles. Uma vez que as vogais acentuadas não sofrem qualquer elevação, surgem numerosos casos em que todas as vogais, excepto a última quando átona, são baixas. Não é pois de estranhar o aparecimento de uma

estratégia geral de não elevação das vogais que se encontram à esquerda do acento principal, estratégia essa que é sistematicamente adoptada em todos os casos em que as terminações apenas ocorrem nesta classe de nomes (ex. *"ax", "ux", "trans", "tur*"). De qualquer modo, o facto de as vogais átonas apresentarem comportamentos diferentes em posição pré e pós-acentual pode ser encarado como um indicador da independência, pelo menos relativa, dos mecanismos que desencadeiam a redução nestas duas posições.

3.4. Leitura e soletração de siglas

Como já foi referido acima, não há uma relação directa entre a forma como são pronunciados os nomes de empresas e de serviços públicos e os processos lexicais utilizados na sua constituição, uma vez que todos eles podem resultar em sequências segmentais idênticas. As siglas propriamente ditas diferem, no entanto, de todos os outros processos, na medida em que nem sempre podem ser oralizadas de acordo com as regras gerais de correspondência grafema-som. Algumas são obrigatoriamente lidas, outras soletradas e outras ainda podem ser oralizadas de qualquer destas formas. Embora pouco frequentes, existem também siglas, cuja oralização é mista, isto é, em que uma parte da sequência segmental é soletrada e a outra parte lida. Decidir quando é que uma sigla (ou parte dela) deve ser lida ou soletrada é um dos problemas fundamentais no tratamento desta classe de nomes.

Na sua versão anterior, o sistema DIXI soletrava todas as siglas constituídas apenas por sequências de consoantes e tentava ler todas as que continham pelo menos uma vogal. A presença de uma vogal é, efectivamente, uma condição necessária para que uma sigla possa ser lida mas não é suficiente para uma escolha adequada do processo de oralização a adoptar: em Acro_Fone, cerca de 4% dos nomes correspondem a siglas que são oralizadas por soletração e cerca de metade destas últimas contêm pelo menos uma vogal. Repare-se, por exemplo, que *AR* (abreviatura de *Assembleia da República*) contém uma vogal, é homógrafa de uma palavra do léxico comum e é, no entanto, sempre soletrada.

A extensão é um factor que deve ser tido em conta: são soletradas todas as siglas com menos de três letras e preferencialmente lidas ou mistas as que têm mais de cinco. Os dois modos básicos de oralização são possíveis com as siglas de extensão intermédia (3 a 4 letras) mas não podem ser utilizados indiscriminadamente. Certos padrões, como os CVCV são sempre lidos (ex.s. *FIFA* [ʼfi.fɐ]; *CEGE* [ʼsɛ.ʒi] etc.) e outros como os VCCC são soletrados (ex.s *APDC* [a.pe.deʼse], *IFPM*; [i.ɛ.fi.peʼɛ.mi]). Com raríssimas excepções, como *SAS,* as siglas CVC são lidas (ex.s *CAP* [ʼkap]; *SIS* [ʼsiʃ]; mas nem todas as que contêm duas vogais, como as VCV ou as CVV o são: (ex.s *IPE* [i.peʼɛ]; *IPO* [i.peʼɔ]; *CEE* [se.ɛʼɛ]).

Observações semelhantes têm sido feitas para outras línguas (cf. [10,15,16]) e estado na origem de tentativas de explicação do modo de oralização das siglas em função da interacção de diferentes factores de ordem prosódica. Assim, por exemplo, para que uma determinada sequência segmental possa ser lida, tem de se prestar a uma análise silábica concordante com o conjunto de princípios gerais e com as restrições específicas da língua, mas tem também de corresponder a um padrão de palavra possível em extensão e em peso. Em determinadas situações, no entanto, as restrições de ordem estrutural e as de peso podem não ser compatíveis e a resolução do conflito depende da sua importância relativa. Plénat (1992) propõe um limiar mínimo e máximo de peso para a oralização das siglas em francês e refere alguns exemplos de conflitos possíveis. O limiar mínimo de duas moras (correspondendo a um monossílabo com rima ramificada ou a um dissílabo), define uma fronteira abaixo da qual uma sigla é obrigatoriamente soletrada e o limiar máximo de três sílabas define outra fronteira, acima do qual ela é obrigatoriamente lida. Estas restrições de peso silábico coexistem com um conjunto de restrições de ordem estrutural que determinam a soletração das siglas cujos constituintes prosódicos podem ser considerados mal-formados como, por exemplo, um pé que contenha um hiato ou que não contenha nenhuma sílaba CV. A proibição do hiato, por exemplo, conduz à soletração das siglas com um padrão CVV uma vez que a forma resultante não ultrapassa o limiar máximo de três sílabas. Quando esse limiar é ultrapassado, o hiato é tolerado e as siglas são preferencialmente lidas. As siglas que admitem os dois modos de oralização serão apenas as que correspondem a casos em que duas restrições contraditórias se equilibram.

A oralização das siglas em Português parece ser, em muitos aspectos, semelhante à que se observa para o francês mas apresenta, naturalmente, algumas discrepâncias significativas que reflectem diferenças de parametrização. O caso mais evidente é, justamente o das siglas com estrutura CVV que, em Português, não são preferencialmente soletradas. Ao contrário do Francês, esta língua admite núcleos ramificados e, por conseguinte, algumas sequências VV são interpretadas como ditongos. É o caso, por exemplo, de *FAO* ou de *JAE* que se pronunciam como [ˈfaw] e [ˈʒaj], respectivamente. As duas vogais da sequência VV podem, no entanto, encontrar-se em hiato, sem que daí resulte necessariamente a soletração da sigla, como acontece com *CIA* [ˈsi.ɐ] , por exemplo. Algumas palavras muito comuns do Português têm uma estrutura exactamente idêntica (ex. *tia* [ˈti.ɐ]; *lia* [ˈli.ɐ]), de qualquer modo, a sequência segmental é muito comum em posição final de palavra, onde o hiato pós-acentual é bem tolerado. Das siglas CVV, apenas são sistematicamente soletradas aquelas em que as duas vogais são idênticas, situação que não ocorre no léxico comum.

O facto de a oralização das siglas poder variar de língua para língua e de essa variação poder ser interpretada em função da sua parametrização específica, para

além do interesse de que se reveste por si próprio, permite pôr em evidência a importância das restrições de ordem estritamente fonológica na aceitabilidade por parte dos falantes de uma dada sequência segmental como “palavra” da língua. O que parece estar em causa não é, contudo, a sua aceitabilidade como “palavra possível” mas como palavra “palavra provável”.

Certas siglas, como *AR*, por exemplo, que são homógrafas de palavras do léxico comum, são certamente palavras possíveis, mas a frequência de ocorrência de monossílabos no léxico é bastante reduzida, se o peso das palavras gramaticais for ignorado (cf. Figura 2). De entre os monossílabos, os que têm ataques vazios são ainda menos frequentes do que os outros**.** Pode reparar-se, com efeito, que o conjunto de factores que contribui para explicar o modo de oralização das siglas também explica a frequência de ocorrência de palavras com a mesma estrutura prosódica no léxico comum.

As línguas em que os núcleos vazios são autorizados podem apresentar sílabas cuja vogal não é realizada foneticamente. Podem também interpretar consoantes não silabificáveis de uma sequência como ataques ou codas de sílabas desse tipo. É por essa razão que podem ser lidas algumas siglas com sequências de obstruintes, de outro modo insilabificáveis. Os constituintes vazios, no entanto, são sempre estruturas marcadas que, de alguma maneira, inibem a leitura das siglas e cujos efeitos parecem ser cumulativos: a inibição de leitura é sempre maior para uma sigla com dois constituintes vazios do que para outra apenas com um. São, por exemplo mais vezes soletradas as siglas em VCv^{0}, como *APE* (e mesmo *IPO*) do que qualquer das outras em VCV ou em CVC. Embora existam núcleos vazios consecutivos em posição final de palavra (por exemplo, em *síntese* [’sĩ.ti.zi] ou *bípede* [’bi.pi.di], normalmente pronunciadas [’sĩ.t.z]e [’bi.p.d], respectivamente) as siglas em CVCC, em que CC são obstruintes, são em geral soletradas (ex. *CEPD*), o mesmo não acontecendo quando CC são silabificáveis e a sigla apenas contém um núcleo vazio (ex. *SERB*, *CELT*) em posição final absoluta.

Com base num pequeno conjunto de regras que dão conta da maior parte destas restrições, foram feitas automaticamente predições sobre o modo de oralização das siglas. A comparação destas predições com as transcrições manuais revelou discordâncias em 5% dos casos. Embora esta taxa de erro possa ser considerada aceitável, subsistem algumas dúvidas àcerca da própria adequação das pronúncias propostas pelos transcritores, uma vez que a maior parte das siglas que ocorrem na base de dados são completamente desconhecidas. Parece conveniente, por conseguinte, testar o desempenho do conjunto de regras sobre um corpus de siglas de utilização corrente, cuja pronúncia não levante dúvidas.

5. Principais resultados e perspectivas futuras

O trabalho que temos vindo a realizar permitiu mostrar que em Português Europeu não há diferenças significativas no modo como são pronunciadas as formas do léxico comum e os nomes próprios. O desempenho do sistema de transcrição fonética automática existente pode ser, por conseguinte, globalmente considerado como bastante satisfatório para estas classes de palavras. Para lidar com os nomes de empresas e serviços públicos, a análise morfológica efectuada pelo sistema de regras tem de ser, no entanto, muito mais elaborada do que aquela que é feita actualmente e tem de recorrer obrigatóriamente a um 'léxico' muito mais extenso.

Em qualquer dos casos, o desempenho da rede neuronal é apenas ligeiramente inferior ao observado para o sistema de regras (3% e 5%, para o léxico comum e os nomes próprios, respectivamente), sendo 44% dos erros cometidos pelos dois métodos rigorosamente idênticos. A análise dos resultados de ambos os métodos permite considerar, no entanto, que muitos dos erros das redes poderão ser evitados se o corpus de treino for melhorado e se a silabação e a atribuição do acento de palavra forem tratados como processos independentes, hipótese que pretendemos vir a testar proximamente.

Alguns aspectos fundamentais para a compreensão do desempenho da rede neuronal não puderam ainda ser explorados, nomeadamente a análise dos seus padrões de activação. Através desta análise será possível saber quais os agrupamentos funcionais que foram feitos e verificar em que medida é que estes coincidem com os preditos pelos modelos linguísticos. Desse ponto de vista, e dado o interesse dos resultados obtidos relativamente à soletração e leitura das siglas, importa analisar também o desempenho das redes sobre o corpus Acro_Fone. Para esse efeito, é necessário, no entanto, recolher um maior número de siglas de utilização comum, cuja pronúncia corrente seja bem conhecida e assegurar um maior equilíbrio na representatividade dos diferentes padrões prosódicos presentes no corpus.

Um dos principais desafios, no contexto do projecto europeu Onomastica, prende-se com a pronúncia nativa de nomes estrangeiros, o que implica considerar diferentes graus de adaptação à estrutura sonora do Português que reflictam diferentes níveis de familiaridade dos falantes com a língua estrangeira. O estudo deste problema é crucial para a utilização generalizada de sistemas de reconhecimento automático e de síntese de fala a partir de texto, nomeadamente para aplicações na área dos serviços de informações telefónicas automáticas. Com base no trabalho realizado no âmbito do projecto Onomastica foi já construído um protótipo de aplicação que permite obter informações de números de telefone a partir do nome do assinante ou moradas a partir dos números de telefone. O

protótipo utiliza o teclado do telefone para a entrada das letras ou dos dígitos, respectivamente. Dado que é possível obter uma cobertura dos nomes e moradas dos assinantes superior a 84% com um conjunto reduzido de nomes, cuja frequência de ocorrência é superior a 100, a presente versão do sistema utiliza um método de concatenação de nomes pré-gravados e soletra as iniciais dos nomes que não estão incluídos na lista. Pretende-se, no entanto, vir assegurar uma maior cobertura dos nomes e incluir siglas e acrónimos, o que obrigará à integração de um módulo de síntese de fala a partir do texto. Para corresponder às necessidades desta aplicação, o sistema DIXI deverá incluir um conjunto de regras sensíveis à categoria dos nomes e terão de ser introduzidas algumas modificações para assegurar o processamento dos compostos graficamente aglutinados, das siglas e dos nomes de origem estrangeira.

AGRADECIMENTOS

Muitas das nossas surpresas e dificuldades foram discutidas com a Amália Andrade e a Alina Villalva a quem gostaríamos de agradecer. Ove Andersen, Paul Dalsgaard e François Yvon facultaram-nos o acesso a trabalhos ainda não publicados, cuja leitura foi também importante para a orientação geral do nosso estudo.

REFERÊNCIAS

[1] Alves, I. M. (1990) - *Neologismo. Criação Lexical*.S. Paulo, Ática.

[2] Andersen, O. e P. Dalsgaard (1994) - "A Self-Learning Approach to the Transcription of the Danish Proper Names". *Proceedings ICSLP 94*, pp.1627-1630.

[3] Andrade, E. e M. C. Viana (1993) - "As sobrodas da translineação". *Actas do 1º Encontro de Processamento da Língua Portuguesa*. Lisboa, Fundação Calouste Gulbenkian.

[4] Andrade, E. e M. C. Viana (1993) - "Sinérese, diérese e estrutura silábica". *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa, Colibri, 1994, pp. 31-42.

[5] Barbosa, J. M. (1965) - *Etudes de Phonologie Portugaise*. Lisboa, Junta de Investigações do Ultramar.

[6] Bisol, L. (1989) - "O ditongo na perspectiva da fonologia atual". *Delta*, 5(2): 185-224.

[7] Gonçalves, F. Rebelo (1947) - *Tratado de Ortografia da Língua Portuguesa*. Coimbra, Atlântida.

[8] Lazzaretto, S. e L. Nebbia (1987) - "SCYLA: Speech Compiler for Your Language". *Proc. of the European Conf. on Speech Tecnology*, Edimburgo, Vol.II, pp. 381-384.

[9] Mateus, M.H.M. A. Andrade, M. C. Viana e A. Villava (1990) - *Fonética, Fonologia e Morfologia do Português*. Lisboa, Universidade Aberta.

[10] McCully C. B. e M. Holmes (1988) - Some notes on the structure of acronyms". *Lingua*, 74(1): 27-43.

[11] Nascimento, F., L. Marques e L. Segura (1987) - *Português Fundamental: Métodos e Documentos*.. Lisboa, INIC-CLUL

[12] Nogueira, R. de Sá (1941) - *Tentativa de Explicação dos Fenómenos Fonéticos em Português*. Lisboa, Livraria Clássica Editora.

[13] Nogueira, R. de Sá (1942) - P Problema da Sílaba. Lisboa, Livraria Clássica Editora.

[14] Oliveira, L., C. Viana e I. Trancoso (1992) - "A rule based text-to-speech system for Portuguese". *Proc. Int. Conf. on Acoustics,Speech and Signal Processing*, San Francisco, Vol II:, pp. 73-76.

[15] Plénat, M. (1991) - "Vers une oralisation des sigles". Iièmes Journées Internationales du GRECO-PRC Communication Homme Machine, EC2 Editeur, Nanterre, pp.363-371.

[16] Plénat, M. (1992) - "Observations sur le mot minimal français". In Laks, B. & Plénat (eds), *De Natura Sonorum*. Saint-Denis, Presses Universitaires de Vincennes, pp 144-172.

[17] Sejnowski, T. J. e C.R. Rosenberg (1987) - "Parallel networks that learn to pronounce English text". *Complex Systems*, 1, pp 145-168.

[18] Trancoso, I., M. C. Viana, F.M. Silva, G. C. Marques e L. C. Oliveira (1994) - "Rule based vs neural network-based approaches to letter-to-phone conversions for Portuguese common and proper names. Proceedings ICSLP 94, pp.1767-1770.

[19] Viana, M.C., E. d'Andrade, L. Oliveira e I.M. Trancoso (1991) -"Ler_PE: um utensílio para o estudo da ortografia do Português". *Actas do VII Encontro da Associação Portuguesa de Linguística*, Lisboa, pp.474-489.

[20] Viana, M.C., E. d'Andrade, L. Oliveira e I.M. Trancoso (1992) -"Uma questão de equilíbrio". *Actas do VIII Encontro da Associação Portuguesa de Linguística, Lisboa*, pp.523-534.

[21] Viana, M.C., I. Trancoso, F.M. Silva, (1994) - "On the pronunciation of proper names and acronyms in European Portuguese". To be presented at the Onomastica Research Colloquium, December 1994, London.

[22] Villalva, A. (1992) - "Compouding in Portuguese". Rivista de Linguistica, 4, pp 201-219.

[23] Villalva, A. (1994) - Estruturas Morfológicas: Unidades e Hierarquias nas Palavras do Português. Dissertação de doutoramento (em preparação).

[24] Vigário, M. e I. Falé (1993) - "A sílaba no Português Fundamental: uma descrição e algumas considerções de ordem teórica". *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa, Colibri, 1994, pp. 465-478.

[25] Yvon, François (1994) - "Règles de Transcription Graphème-Phonème pour la Prononciation Automatique des Sigles" . *Lynx*, 30 (no prelo).