Léxicos de Pronúncia:
A Experiência do Projecto ONOMASTICA

Isabel Trancoso M. Céu Viana Isabel Mascarenhas

INESC / IST CLUL CLUL

INESC, R. Alves Redol, 9, 1000 Lisboa

Resumo

Esta comunicação aborda alguns aspectos relevantes da pronúncia de nomes próprios, tentando simultaneamente fazer um apanhado da experiência da equipa de investigadores portugueses que participou no projecto europeu ONOMASTICA.

O principal resultado do projecto foi a constituição de dois recursos linguísticos de importância fundamental: o conjunto dos 11 léxicos de pronúncia nacionais e o léxico de pronúncia inter-língua. O artigo começa assim por descrever os objectivos do consórcio, e apresentar o formato e o conteúdo dos dois tipos de léxico. Em seguida, far-se-á uma breve abordagem do problema da conversão automática grafema-fone, cuja aplicação é praticamente obrigatória sempre que se pretenda desenvolver léxicos de pronúncia desta dimensão, e descreve comparativamente as várias metodologias adoptadas durante o projecto.

Dois dos aspectos mais importantes do processamento dos dois tipos de léxico são o da pronúncia dos acrónimos e o da nativização da pronúncia de nomes estrangeiros. Esta comunicação aborda sobretudo o último, dado que constitutuiu o tema principal de investigação na parte final do projecto.

1 INTRODUÇÃO

O projecto Europeu ONOMASTICA, recentemente concluído, foi uma iniciativa de investigação à escala europeia, no âmbito do programa LRE (Linguistic Research and Engineering) que tinha como objectivo a construção de um léxico multilingue de pronúncia de nomes próprios, em que foram consideradas 11 línguas europeias [8]. No consórcio, coordenado pelo CCIR, participou por cada país um parceiro académicos e um associado:

- Alemão: Inst. für Fernmeldetechnik, Berlim + Deutche B. Telekom
- Dinamarquês: CPK, Univ. Aalborg + Jydsk Telefon
- Espanhol: UPM, Madrid + Telefónica
- Francês: ENST, Paris + France Telecom
- Grego: Dept. Electrotechnical Engineering, Univ. Patras + Intrasoft
- Inglês: CCIR, Univ. Edinburgh + BT Laboratories
- Italiano: Inst. de Ling. Computacional, Pisa + CSELT
- Neerlandês: Dept. Language and Speech, Nijmegen + PTT Research
- Norueguês: SINTEF DELAB, Trondheim + Norwegian Telecom Research
- Português: INESC, Lisboa (com a colaboração do CLUL) + Portugal Telecom (inicialmente TLP)
- Sueco: Kungl Tekniska Hogsk., Estocolmo + Telia (Infovox).

Os parceiros associados, em grande maioria constituídos por operadoras telefónicas, forneceram os ficheiros de dados incluindo nomes de pessoas, cidades, ruas e companhias.

A grande motivação para um projecto com estes objectivos reside no facto de, em geral, o desempenho dos sistemas de conversão grafema-fone para nomes próprios ser 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 para o léxico comum e que só raramente têm sido contempladas peculiaridades da pronúncia dos nomes próprios que podem obedecer a regras morfo-fonológicas bastante diferentes.

Parte do problema reside na mobilidade dos nomes próprios, que "viajam" com as pessoas de um país para outro, mostrando diferentes graus de adaptação à estrutura sonora da língua

do país de acolhimento. Existem, no entanto, outras fontes de problemas. Por um lado, a ortografia dos nomes próprios nativos pode ser bastante conservadora, apresentando distribuições de grafemas que deixaram de ser contempladas e que, naturalmente, se prestam a interpretações fonéticas incorrectas. Por outro lado, a pronúncia de nomes de companhias levanta também sérios problemas, uma vez que os acrónimos podem obedecer a regras bastante distintas das regras gerais observadas para o léxico comum.

Um dos objectivos fundamentais do projecto era a construção de dicionários de pronúncia para cerca de um milhão de nomes por língua, o que não seria viável a não ser de uma forma semi-automática. Uma parte do esforço de investigação inicial foi consequentemente direccionada para a melhoria dos sistemas de regras, já existentes para a maior parte das línguas, de forma a lidar com os problemas particulares levantados pela pronúncia de nomes próprios. Neste âmbito, dispendeu-se um esforço muito significativo no desenvolvimento de métodos de auto-aprendizagem de conversão grafema-fone e na sua comparação com os sistemas de regras. Um segundo tópico de investigação particularmente interessante neste projecto foi a pronúncia de acrónimos, estudada com bastante ênfase por dois dos parceiros. A parte final do projecto foi dedicada ao estudo dos problemas de nativização da pronúncia de nomes estrangeiros, tendo-se constituído para isso uma matriz de nomes de cada um dos países com as respectivas pronúncias nativizadas em todas as línguas.

Esta comunicação aborda todos os assuntos acima mencionados, começando por descrever o conteúdo e formato dos léxicos de pronúncia construídos: os 11 léxicos nacionais e o léxico inter-língua. Segue-se-lhe uma descrição muito abreviada dos métodos de conversão grafema-fone do tipo auto-aprendizagem, uma discussão dos problemas levantados pela pronúncia de acrónimos e, por último, um levantamento dos factores que podem influenciar a nativização da pronúncia de nomes estrangeiros.

Note-se que alguns dos temas focados nesta comunicação (conversão grafema-fone por meio de redes neuronais e pronúncia de acrónimos) foram já tratados de uma forma mais detalhada num artigo apresentado no Congresso Internacional sobre o Português [12]. A sua repetição na presente comunicação, embora de uma forma muito mais abreviada (secções 3.1 e 4, respectivamente), justifica-se apenas pelo desejo de reunir num único documento os principais resultados da experiência acumulada ao longo do referido projecto. Pelo facto de a parte final do projecto ter sido dedicada aos problemas de nativização levantados pelo léxico inter-língua, esta será a secção relativamente mais detalhada da presente comunicação.

2 OS LÉXICOS DE PRONÚNCIA DO PROJECTO ONOMASTICA

2.1 Os 11 léxicos nacionais

O número de entradas do léxico ONOMASTICA difere significativamente de língua para língua, variando desde cerca de cem mil até perto de um milhão e totalizando 8,5 milhões de nomes. Uma justificação óbvia para esta discrepância é a diferença entre as populações dos vários países. Outra justificação, que importa também mencionar, é o facto de alguns parceiros tartarem apenas entradas constituídas por palavras isoladas, enquanto que outros consideraram também entradas compostas (i.e., *St. Paul's Cathedral* pode ser considerada como uma única entrada).

Todas as entradas foram automaticamente processadas de modo a obter transcrições fonéticas largas e uma percentagem significativa foi depois verificada manualmente por um foneticista, tendo sido admitidas, no máximo, 5 transcrições alternativas para cada entrada, etiquetada com a respectiva categoria (nome de baptismo, sobrenome, nome de companhia, nome de rua e nome de cidade ou região). Para algumas línguas estava também disponível informação sobre frequência de ocorrência e etimologia. A Tabela 1 ilustra a estrutura de cada entrada do léxico.

A certificação da qualidade das transcrições foi uma preocupação constante ao longo do projecto. Definiram-se assim inicialmente três bandas: banda I, na qual se incluem as transcrições verificadas pelo menos por um transcritor que tem a certeza da sua correcção; banda II, na qual se incluem as transcrições sobre as quais susbsistem algumas dúvidas por parte do transcritor que as verifica; e banda III, na qual se incluem as transcrições não verificadas.

A qualidade das transcrições foi posteriormente avaliada por um auditor independente para cada língua que verificou 1000 nomes seleccionados aleatoriamente em cada banda.

Para as línguas em que se dispunha de informação sobre frequência de ocorrência, realizaram-se também estudos sobre a distribuição estatística de nomes e a cobertura correspondente. Como exemplo, tome-se o caso do léxico correspondente aos nomes e moradas dos habitantes das duas maiores cidades de Portugal. Das cerca de 100.000 entradas diferentes deste léxico, aproximadamente metade correspondem a ocorrências únicas (tipicamente pertencentes a pequenas companhias, nomes estrangeiros e erros de ortografia). Apenas 3% das entradas ocorrem mais de 100 vezes e 13% mais de 10 vezes. Com este último subconjunto reduzido de entradas consegue-se no entanto uma cobertura de cerca de 84% de todos os nomes completos do directório nacional. A representatividade do léxico ONOMASTICA é, portanto, muito elevada.

CAMPO	ABREVIATURA
PRINCÍPIO DO REGISTO	SOO:
FIM DO REGISTO	EOO:
IDENTIFICADOR	ENT:XX123456
NOME (ORTOGRAFIA)	LBO:
COMENTÁRIOS (TEMPORÁRIOS)	CMT:
TRANSCRIÇÕES FONÉTICAS (MAX.5)	XXn:
BANDA DE QUALIDADE	QUi:
TRANSCRITOR	WHn:
ETIMOLOGIA	ETn:
FREQUÊNCIA	FQn:
ANOTAÇÕES (PERMANENTES)	ANn:
CATEGORIA	CTn:

Tabela 1: Estrutura de cada entrada do léxico. XX - Código do país. n - Número da transcrição (n=1,2,3,4,5). i - Número da banda de qualidade (i=1,2,3)

2.2 O léxico de pronúncia inter-língua

Enquanto que o conjunto dos 11 léxicos de pronúncia nacionais pode ser de imediato explorado comercialmente, em particular no desenvolvimento de aplicações no domínio das telecomunicações, o léxico inter-língua deve ser encarado antes como ferramenta de investigação para o estudo de pronúncias "nativizadas". Está limitado a 1000 nomes por língua, trocados entre os 11 parceiros, dando assim origem a uma matriz de 11 pronúncias nativizadas para cada nome estrangeiro (11 x 11 x 1000).

O critério adoptado na selecção destes nomes foi o de salientar o potencial deste tipo de dicionários em aplicações multilingues de reconheciemento de fala envolvendo utilizadores em diversos países europeus. Seleccionaram-se assim nomes de cidades, aeroportos, estações, monumentos e outros sítios de interesse cuja dimensão, significado histórico ou importância geográfica (nomeadamente em termos de transportes) justificam a sua inclusão em guias turísticos. As aplicações em vista são as mais provavelmente utilizáveis por falantes não-nativos, implicando assim o reconhecimento de pronúncias consideravelmente distintas: informação sobre viagens, reserva de bilhetes, sistemas de navegação, informação sobre o estado das estradas, previsões meteorológicas, etc.

A Tabela 2 especifica o conteúdo da matriz inter-língua trocada entre os parceiros em termos de categorias. A primeira inclui nomes de cidades, províncias, regiões e ilhas. Para algumas línguas, esta é a única categoria presente. Para outras, adoptaram-se vários critérios de modo a restringir o número de cidades seleccionadas (dimensão, importância do ponto de vista administrativo ou turístico, ou ainda associação a produtos famosos tais como vinhos ou queijos). A segunda categoria inclui nomes de rios, lagos, baías, canais, montanhas, vulcões, cabos, golfos, cavernas e outros acidentes geográficos. Há obviamente alguma sobreposição entre estas primeiras categorias, dado que, por exemplo, os nomes de rios mais importantes, são frequentemente também nomes de regiões. A terceira categoria inclui nomes de igrejas, museus, pontes, torres, palácios, termas e outros sítios de interesse turístico. Para as cidades mais importantes, incluem-se também nomes associados a zonas relevantes (p.e., estações de comboio, avenidas, praças, parques, etc., os quais são frequentemente associados ao monumento mais próximo). Para algumas línguas, de origem românica sobretudo, uma grande percentagem de nomes desta categoria tem cariz religioso. Quadros famosos e outros tesouros de arte podem também ser incluídos. A última categoria inclui uma miscelânea de informação: festas de interesse turístico, gastronomia, nomes de cidades estrangeiras, etc.

CATEGORIA	I	II	III	IV
de	100	0	0	0
dk	√	√	√	√
es	26	43	22	8
fr	91	<1	8	2
gr	100	0	0	0
it	√	√	√	√
nl	74	4	22	0
nw	√	√	√	√
pt	58	2	40	0
se	√	√	√	√
uk	√	√	√	√

Tabela 2: Percentagem de entradas em cada categoria para as 11 línguas. √ - percentagens não disponíveis

2.3 O CD-ROM ONOMASTICA

Uma parte significativa dos léxicos de pronúncia construídos no âmbito do projecto ONOMASTICA está já incluída num primeiro CD-ROM que comporta presentemente 25.000 entradas (banda I) de 8 línguas, para além do léxico inter-língua.

Por uma questão de uniformidade entre todas as línguas, optou-se por não incluir informação sobre frequência de ocorrência e etimologia. Embora cada um dos parceiros tenha adoptado a sua própria versão computacional do alfabeto fonético para transcrever quer o seu léxico nacional quer o léxico inter-língua, as transcrições fonéticas incluídas no CD-ROM foram traduzidas para o "International Phonetics Association Standard Computer Coding" [2].

O projecto estipulou a utilização de transcrições fonéticas largas, não incluindo necessariamente estrutura prosódica. No entanto, admitiu-se a utilização opcional de transcrições mais estreitas incluindo, por exemplo, fenómenos de lenição. Note-se também que a existência de entradas compostas por mais de uma palavra (muito frequentes também no léxico inter-língua) implica a utilização de regras de sandhi externo (e.g., *Aix en Provence*).

De modo a fornecer uma interface adequada para aceder aos dados armazenados em CD-ROM, foi também desenvolvido um API (Application Programmers' Interface) pelo CCIR. Escrito em C, este API pode ser utilizado quer em DOS quer em Windows, oferecendo as funções típicas de abertura, busca, leitura e fecho de ficheiros de dados.

Para demonstrar o uso de chamadas a estas funções, foi também desenvolvido um programa em Visual Basic. A Fig. 1 ilustra a sua utilização.

3 CONVERSÃO AUTOMÁTICA GRAFEMA-FONE

Embora a maioria dos parceiros dispusesse já de conjuntos de regras para a conversão automática grafema-fone no início do projecto, regras essas que houve que modificar de modo a contemplarem algumas especificidades dos nomes próprios, o esforço de investigação nesta área concentrou-se no desenvolvimento e teste de métodos de auto-aprendizagem. Entre estes, salientam-se os métodos baseados em redes neuronais, tanto convencionais (retro-propagação), como auto-organizáveis, e outros métodos de aprendizagem simbólica que vão desde a busca de tabelas à aprendizagem por analogia.

O trabalho de melhoramento das regras para poderem lidar não só com o léxico comum mas também com nomes próprios variou muito de língua para língua. Para o Português, por exemplo,

Figura 1: Exemplo de utilização do programa em Visual Basic.

testes comparativos efectuados com o sistema de regras sobre um subconjunto do léxico comum, com cerca de 8.000 palavras e sobre um corpus constituído pelos 15.000 nomes próprios mais frequentes (excluindo acrónimos), resultaram em percentagens de palavras erradamente transcritas muito semelhantes (5% vs. 7%, respectivamente). Isto prova que em Português, ao contrário do que é muitas vezes referido para outras línguas, as tabelas de correspondência grafema-fone, não diferem fundamentalmente para os dois corpora, em termos absolutos. De notar, no entanto, que o subcorpus de nomes próprios inclui apenas os nomes mais frequentes e que são necessárias algumas pequenas modificações das regras para ter em conta algumas características de nomes próprios que não se encontram nos dois corpora de teste: a ocorrência de algumas consoantes germinadas (*tt*, *ll*, *mm*, etc.) em nomes de família de origem estrangeira ou com ortografia antiga e a ocorrência de sequências de grafemas pouco comuns em acrónimos e nomes estrangeiros. Trataremos destes dois problemas particulares em capítulos separados.

3.1 Redes neuronais

A aplicação de redes neuronais à conversão grafema-fone data de 1987, quando o sistema NETTALK foi apresentado pela primeira vez [9]. Tal como neste trabalho pioneiro, o tipo de rede neuronal utilizado pela equipa Portuguesa foi do tipo rede multi-camada convencional, treinada pelo algoritmo de retro-propagação de erros [10]. A fase de aprendizagem foi precedida por um alinhamento automático entre as cadeias de grafemas e respectivas transcrições do corpus de

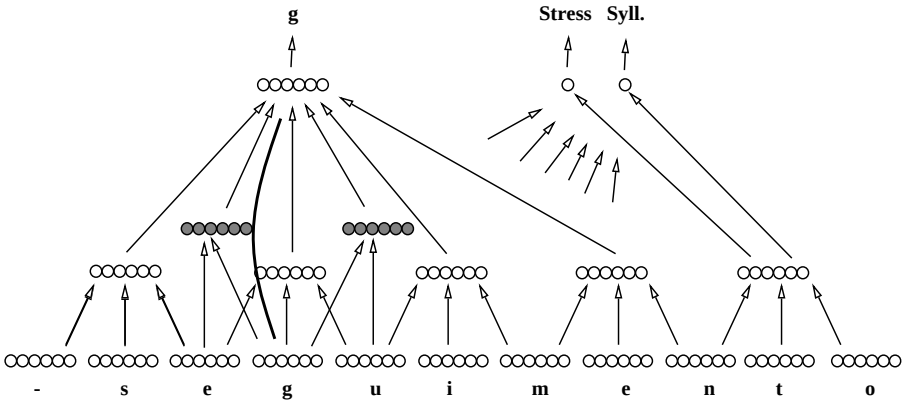

Figura 2: Arquitectura da rede neuronal multi-camada

treino. Tornou-se conveniente 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 treino da rede foi 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.

Testaram-se várias arquitecturas e comprimentos de contexto. A rede ilustrada na figura 2 tem uma camada de entrada constituída por 11 grafemas: o grafema a transcrever, 3 grafemas à sua esquerda e 7 à sua direita, dos quais apenas 5 são utilizados para a transcrição fonética, sendo os últimos 2 utilizados 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). 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. Existem 47 unidades de saída, uma por cada uma das 45 unidades fonéticas consideradas (incluindo unidades simples e complexas), uma para a marca de acento principal e outra para a marca de fronteira de sílaba. De modo a diminuir o número de pesos a ajustar, adoptaram-se pesos partilhados. Existe ainda uma ligação directa da entrada para a saída.

A rede foi treinada com um subconjunto aleatoriamente seleccionado do léxico comum, disjunto do de teste, contendo cerca de 100.000 fones. Ao fim de 8 iterações, o erro ao nível de segmento era já de $1,5\%$, baixando para 1% ao cabo de 40 iterações. Testes com os dois cor-

pora acima mencionados resultaram numa percentagem de erros, a nível de palavras, ligeiramente mais baixa para o léxico comum do que para o léxico de nomes próprios (7% vs. 12%). É interessante constatar que uma percentagem significativa das palavras em que as regras falham é também transcrita erroneamente pela rede neuronal (74% dos casos, para nomes próprios), e que aproximadamente em metade destas palavras, os erros são idênticos em ambas as abordagens.

A maioria dos erros diz respeito à transcrição dos grafemas *e*, *o* e *x*. A rede, contudo, tem maior dificuldade em lidar com a nasalidade, a redução vocálica e a ditongação. O primeiro tipo de dificuldade é evitado no sistema de regras através da colocação de marcas de silabificação antes da fase de transcrição fonética. O segundo, por seu lado, é evitado também através da colocação de marcas de acento antes da transcrição. É de notar, no entanto, que muitos dos problemas de generalização se devem à fraca representatividade das sequências de grafemas correspondentes no corpus de treino.

Saliente-se por último que muitos aspectos interessantes do desempenho das redes neuronais estão ainda por explorar. Em particular, a análise dos padrões de activação de modo a determinar agrupamentos funcionais.

3.2 Busca em tabelas

A abordagem baseada em busca de tabelas foi desenvolvida pelo parceiro académico Dinamarquês (CPK). O pacote de software fornecido a todos os parceiros para teste nas diferentes línguas designa-se por SELEGRAPH [1]. A principal diferença entre este tipo de abordagem de auto-aprendizagem e a descrita anteriormente reside na falta de capacidade de generalização da busca em tabelas, desvantagem essa que até certo ponto é contrabalançada pela maior rapidez do processo de treino. As abordagens deste tipo são treinadas com base em pares de cadeias grafema-fone devidamente alinhadas, determinando-se dinamicamente quais os contextos esquerdo e direito mínimos que permitem mapear cada um dos grafemas no fone correcto com uma certeza absoluta.

O processo de treino propriamente dito é assim precedido por duas fases: o alinhamento, tal como para a rede neuronal, e o cálculo da informação mútua, que determina para cada grafema, quantos grafemas contextuais há que incluir e a ordem em que devem ser considerados.

O treino resulta numa estrutura em árvore em que cada nó armazena para um dado grafema num dado contexto, estatísticas do número de vezes em que ocorreu cada fone possível. Utilizam-se mapeamentos por defeito para conversões grafema-fone ambíguas e para palavras desconhecidas que contenham sequências de grafemas não contempladas no corpus de treino.

Tal como para a rede neuronal, a abordagem baseada em busca de tabelas foi treinada para o

Português com um subconjunto do léxico comum, e testada tanto para um subconjunto disjunto do mesmo léxico como para um subconjunto do léxico de nomes próprios. Uma análise dos erros revela o mesmo tipo de dificuldades que as encontradas pela rede, embora com uma maior frequência, o que evidencia a sua falta de capacidade de generalização [10].

Estes dois métodos de auto-aprendizagem mostraram que para o Português, podem ter potencialmente um desempenho tão bom como o sistema de regras, desde que se utilize para treino um corpus de grande dimensão e / ou se adoptem estratégias separadas para a colocação de marcas de silabificação e atribuição de acento.

3.3 Abordagens baseadas em analogia

As abordagens baseadas em analogia foram desenvolvidas nomeadamente pelos parceiros académicos Franceses e Italianos. Este tipo de abordagens também assume para o processo de aprendizagem um corpus de treino com as ortografias e transcrições devidamente alinhadas. Para cada palavra de teste, a pronúncia é determinada por analogia com pronúncias conhecidas de palavras "parecidas", através da aplicação de duas funções: uma função de mapeamento, definida sobre cadeias de símbolos ortográficos, e uma função de recombinação, definida sobre cadeias de símbolos fonéticos. A primeira projecta a ortografia da palavra de teste sobre as palavras de treino de modo a seleccionar as subcadeias mais "análogas". A função de recombinação vai então "juntar" as transcrições correspondentes às subcadeias assim seleccionadas.

Um dos critérios mais óbvios à primeira vista para realizar esta selecção consiste em procurar as subcadeias ortográficas de treino mais longas que coincidem com a de teste quer começando no extremo esquerdo desta quer no extremo direito. A referência [6], no entanto, demonstra com exemplos muito simples, a inadequação deste critério simplista de maximização do comprimento, defendendo a estruturação do corpus de treino em famílias paradigmaticamente relacionadas e a adopção de um critério que selecciona as subcadeias mais longas que satisfaçam duas condições: (1) pertençam à mesma família e (2) partilhem uma parte central que impeça a formação de sequências fonotácticas indesejáveis.

É sabido que as crianças aprendem a ler através de um raciocínio de alguma forma baseado em analogia. No entanto, este tipo de abordagens não foi ainda devidamente explorado, existindo uma vasta área de investigação sobre o papel dos vários factores em jogo e sobre como encontrar funções de mapeamento e recombinação flexíveis e simultaneamente manejáveis do ponto de vista computacional. Apesar destas dificuldades, os resultados citados para nomes Italianos e Ingleses ultrapassam os obtidos com abordagens baseadas em regras ou em procedimentos de busca em

tabela. A tarefa, contudo, constituiu um maior desafio para esta última língua.

4 ACRÓNIMOS

Para algumas das línguas tratadas no projecto, os dados fornecidos pelos parceiros associados continham também nomes de companhias. A pronúncia de acrónimos, que constituem uma parte muito significativa do conjunto dos nomes de companhias, foi um dos tópicos de investigação estudados em particular pelas equipas Francesa [14] e Portuguesa [12].

Os acrónimos constituem cerca de 38% das 50.000 entradas mais frequentes do ficheiro de dados nacional. Para esta categoria de nomes, tanto o desempenho do sistema de regras como o da rede neuronal se revelaram claramente insatisfatórios (apenas 57% e 49% de resultados coincidentes com as transcrições fonéticas manuais, respectivamente). Para além disso, a sua pronúncia por falantes nativos varia consideravelmente. Estes dois factos motivaram assim o estudo dos processos lexicais utilizados na formação de acrónimos e da sua relação com a variabilidade observada que relataremos nas duas subsecções seguintes. A subsecção final é dedicada ao estudo de uma classe especial de acrónimos que merece um ênfase especial - as siglas.

4.1 Processos lexicais de formação de acrónimos

Na constituição de nomes de companhias, são utilizados modificadores morfológicos, radicais e palavras do léxico comum, primeiros nomes, apelidos, topónimos e praticamente todas as abreviaturas destes, combinadas quer entre si quer com palavras estrangeiras ou com terminações características desta classe de nomes.

Podem-se encontrar na sua formação vários tipos de processos de criação lexical comuns em Português: acronímia (em sentido restricto), amálgama, 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 seleccionados (amálgama). A distinção fundamental não está propriamente no número de letras que são retidas mas nos critérios que presidem à sua selecção: enquanto os acrónimos são sempre construídos para serem "lidos", as siglas podem ser lidas ou soletradas, justificando-se algumas apenas pela facilidade da escrita.

Uma grande percentagem dos acrónimos encontrados no léxico Português, contudo, resulta de um tipo de processos de formação lexical não muito frequente no nosso léxico comum - a

composição. A principal distinção entre compostos do léxico comum, de acordo com [13], é entre compostos de palavras e compostos de radicais. Os primeiros podem ter tantas vogais abertas quantos os elementos constituintes, enquanto que os segundos têm, para além dessas, uma vogal de ligação, /i/ ou /ɔ/ que, no segundo caso, não sofre também elevação. Graficamente, os dois tipos são distinguíveis pelo facto de que os primeiros se escrevem tipicamente como palavras separadas (frequentemente com hífenes), e os segundos se escrevem como uma única palavra. Por conseguinte, os compostos de palavras não necessitam de qualquer tratamento especial para serem correctamente transcritos e a maior parte dos compostos de radicais é identificável com base numa lista relativamente reduzida de morfemas presos, na sua maioria de origem greco-latina. Para os acrónimos, no entanto, 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.

4.2 A pronúncia de acrónimos

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: (1) directamente junto de um conjunto de empresas para averiguar qual a origem e pronúncia dos seus nomes; (2) junto de 10 falantes de formação escolar de nível universitário, a quem foi pedida a leitura de uma lista de 100 itens, aleatoriamnete seleccionados e não anunciados na comunicação social. O contacto directo com as empresas mostrou, sobretudo, a grande variedade de critérios que podem presidir à escolha de um nome: 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, etc.. A pronúncia pretendida, contudo, nem sempre é a mais frequentemente adoptada pelos falantes nativos de Português. É comum encontrarem-se múltiplas pronúncias aceitáveis para um único acrónimo e, nesses casos, nem sempre é fácil distinguir a mais provável, sendo no entanto simples detectar as pronúncias claramente inaceitáveis. O facto de, no teste de leitura acima mencionado, 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, como *globomar*, por exemplo. Dado que a vogal de ligação "o" em compostos de radicais é graficamente idêntica à marca do masculino dos compostos de palavras, as formas deste tipo são inerentemente

ambíguas e prestam-se a oscilações de pronúncia. *Globomar* foi pronunciada como [glo.bɔ'mar] por 40% dos falantes e como [glo.bu'mar] por 60%. Contudo, o reconhecimento de palavras ou de radicais dentro de palvras gráficas não parece ser uma tarefa que faça parte dos hábitos de leitura dos Portugueses. Se fizesse, então uma forma como *alfasom* seria invariavelmente tratada como composto de palavras (*alfa + som*) e pronunciada como [aɬ.fɐ'sõ], como pretendido pelos seus criadores. Contudo, em 60% dos casos, esta forma é tratada como uma palavra simples e o "s" em posição intervocálica é pronunciado como [z],seguindo as regras gerais de pronúncia. As regras de correspondência grafema-fone parecem assim prevalecer sobre a análise morfológica na tarefa de leitura.

A maior parte das pessoas tem consciência de que os nomes de empresas e 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 é 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 (e.g., *-ax, -ux*).

4.3 Leitura e soletração de siglas

Como já foi acima mencionado, as siglas levantam problemas específicos de pronúncia. 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 é soletrada e a outra parte lida. Decidir quando é que uma sigla deve ser lida ou soletrada é um dos problemas fundamentais no tratamento desta classe de nomes.

Na sua versão anterior, o nosso sistema de regras soletrava todas as siglas constituídas apenas por sequências de consoantes e tentava ler todas as que continham pelo menos uma vogal. Esta última condição é, de facto, uma condição necessária para que uma sigla possa ser lida, mas não é suficiente: no nosso léxico, cerca de metade das siglas que são oralizadas por soletração contêm pelo menos uma vogal. A extensão é um factor que deve ser tido em conta: à parte raras excepções, sao soletradas todas as siglas com menos de 3 letras e preferencialmente lidas ou mistas as que têm mais de 5. 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 (e.g. *FIFA*) e outros como as VCCC são soletrados (e.g. *APDC*). Com raríssimas excepções, as siglas CVC são lidas (e.g. *CAP*); contudo, nem todas as que contêm duas vogais como as VCV ou CVV o são (e.g. *IPE*).

Observações semelhantes têm sido feitas para outras línguas e estado na origem de tentativas de explicação do modo de oralização das siglas em função da interacção de diferentes restrições. Plénat [7] propõe um limiar mínimo e máximo de peso para a oralização das siglas em Francês. 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 da qual ela é obrigatoriamente lida. Estas restrições de peso silábico, contudo, interactuam com outras restrições prosódicas. As siglas CVV, por exemplo, são geralmente soletradas em Francês mesmo quando acima do limiar mínimo, o que pode ser explicado pela necessidade de evitar o hiato [7]. A proibição do hiato, contudo, pode ser ultrapassada pela extensão.

Embora um raciocínio semelhante se possa fazer para ter em conta a pronúncia de siglas em diversas línguas, o seu modo de oralização difere frequentemente. Em Português Europeu (EP), por exemplo, onde a proibição do hiato é também uma restrição muito forte, as siglas CVV são preferencialmente lidas. Uma explicação possível para esta diferença está no facto desta língua admitir núcleos ramificados e o hiato poder ser evitado por ditonguização. Quando isto não for possível, devido ao acento na vogal alta, as siglas também não são soletradas. A sigla *CIA*, por exemplo, é mais frequentemente lida dado que, no caso contrário, implicaria duas violações da restrição do hiato. A leitura também está mais de acordo com o padrão de acento do Português. O estudo da oralização das siglas em várias línguas reflecte, portanto, diferenças de parametrização.

A interacção entre restrições está também relacionada com a probabilidade de ocorrência de padrões de palavras na língua. Certas siglas tais como *AR* (abreviatura de *Assembleia da República*), por exemplo, que é homógrafa de uma palavra do léxico comum, são obviamente palavras possíveis, embora como nomes de companhias ou serviços públicos são sempre soletradas. De facto, a frequência de ocorrência de palavras monossilábicas no léxico é muito reduzida (se ignorarmos o peso das palavras funcionais) e as que têm um ataque vazio são ainda menos frequentes que as outras.

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 presentes no léxico. Em 95% dos casos, as predições concordaram com as opções escolhidas pelos transcritores manuais.

5 NATIVIZAÇÃO DA PRONÚNCIA DE NOMES ESTRANGEIROS

Um dos aspectos mais interessantes do processamento da matriz inter-língua consistiu na definição de pronúncias "nativizadas" de nomes estrangeiros [11]. Existe uma grande variedade de critérios que podem presidir a esta definição. A pronúncia "nativizada" por defeito adoptada pelo consórcio é a de um falante nativo relativamente pouco exposto no passado a línguas estrangeiras. Em termos gerais, esta pronúncia por defeito segue de perto a transcrição gerada pelas regras de conversão grafema-fone da língua nativa. O conjunto de regras, no entanto, tem que ser alargado de modo a ter em conta diacríticos inexistentes na língua nativa e sequências de grafemas pouco familiares.

5.1 Factores que influenciam a nativização

Entre esta transcrição nativizada por defeito e a transcrição correspondente à língua original do nome, existe um grande leque de pronúncias possíveis, pelo que o consórcio previu a inclusão opcional de transcrições adicionais que reflectissem uma exposição crescente a línguas estrangeiras. Este leque de transcrições resulta da conjugação de vários factores. Um dos factores é a capacidade do leitor para identificar um dado nome como estrangeiro a partir da sua ortografia. De facto, muitos nomes estrangeiros podem não ser identificados como tal quando a sua ortografia estiver de acordo com as restrições fonotácticas da língua nativa. Por outro lado, algumas sequências de grafemas inexistentes na língua podem dar origem à identificação do nome como estrangeiro, mas não à identificação correcta da sua origem. Aliás, este é um dos aspectos em que o léxico inter-língua difere dos léxicos nacionais, dado que, nestes últimos, a etimologia de um nome não é geralmente conhecida e a tarefa de a adivinhar é deixada ao cuidado do transcritor.

Mesmo admitindo que a origem de um nome é reconhecida correctamente, há muitos outros factores que podem influenciar o grau de nativização: o conhecimento das regras de pronúncia da língua estrangeira, o conhecimento da pronúncia local do próprio nome e a capacidade de pronunciar os sons da língua estrangeira. Os dois primeiros factores estão associados àquilo que, neste contexto, designaremos como "competência de leitura" e o último à "competência de pronúncia" do falante [5]. A competência de leitura depende do grau de afinidade entre a língua estrangeira e a língua nativa (por exemplo, se pertencem ao mesmo grupo de línguas Românicas ou Germânicas) e da familiaridade que o falante tem com a língua estrangeira. Esta é tipicamente superior para línguas como o Inglês e (talvez em menor escala) o Francês, dado que são ensinadas

na escola secundária em muitos países europeus. Sempre que um falante ignora por completo as regras de pronúncia da língua estrangeira, procura tipicamente padrões semelhantes nas línguas que conhece de modo a escolher a sua pronúncia.

A combinação de diferentes graus de competência de leitura e de pronúncia dá origem a um vasto leque de pronúncias possíveis, tal como acima mencionámos. Teoricamente, contudo, tem interesse definir um falante nativo hipotético que conhece perfeitamente as regras de leitura da língua estrangeira, mas está restricto ao conjunto de fones da sua língua nativa. A pronúncia deste hipotéctico falante foi fornecida por alguns parceiros para algumas das línguas. A comparação entre as várias pronúncias nativizadas de cada nome nas diversas línguas está actualmente em curso. Daqui se espera poder tirar alguns dados interessantes sobre as afinidades entre as línguas nativa e estrangeira.

Também potencialmente interessante é a comparação entre esta segunda pronúncia nativizada, assumindo plena competência de leitura, e a pronúncia nativizada por defeito que assume competência nula quer em termos de leitura quer de pronúncia. Esta comparação foi feita para um conjunto reduzido de 250 nomes de cinco línguas seleccionadas de modo a reflectir diferentes graus de familiaridade e afinidade com a língua nativa (Neerlandês, neste estudo particular [4]). As línguas familiares eram o Alemão, o Francês e o Inglês, todos leccionados na escola, e as não familiares o Sueco e o Italiano. As pronúncias nativizadas por defeito foram geradas por regra e posteriormente comparadas com a nativizada "ideal". As transcrições foram alinhadas por um algoritmo de programação dinâmica de modo a procurar a melhor concordância entre cadeias de símbolos fonéticos. O algoritmo permitiu calcular medidas de distância acumulada mínima que foram posteriormente submetidos a uma análise de variância. O Inglês e o Francês atingiram medidas de distância superiores a 1 (1.4 e 1.7, respectivamente), o que significa que as regras de conversão grafema-fone para o Neerlandês geram transcrições bastante diferentes das correspondentes à pronúncia nativizada "ideal" para estas duas línguas. Para o Sueco e o Italiano, a concordância é melhor (0.6 e 0.7, respectivamente). Os melhores resultados foram obtidos para o Alemão (0.4). Isto parece sugerir que a afinidade entre a língua estrangeira e a língua nativa desempenha um papel primordial.

5.2 Inventários sonoros

A competência de pronúncia, neste contexto, diz respeito à capacidade de pronunciar sons que não existem na língua nativa do falante. Muitos desses sons são pura e simplesmente aproximados por sons nativos. As vogais nasais francesas, por exemplo, são representadas em Norueguês, onde

não existem, por uma vogal seguida de uma consoante nasal; os sons [œ] e [ø] são tipicamente substituídos por [ɐ] em Português, etc.

Em muitos casos, no entanto, o conjunto de fones da língua nativa é alargado de modo a incorporar fones de outras línguas. Em Italiano, por exemplo, ao alfabeto fonético original foram adicionados 5 símbolos novos: [ʒ] (para transcrever o *j* Francês, como em *journal*); [h] (para o *j* Espanhol, tal como em *Julio*; [y] (para o *u* Francês, como em *Durand*); [œ] (para a primeira vogal de *Voeller* em Alemão); e o schwa [ə] para imitar alguns sons estrangeiros (e.g. *de* em Francês), e como vogal muda a inserir sempre que necessário para pronunciar sílabas de outro modo impronunciáveis, como na sílaba final de *Argenteuil* em Francês.

O facto de certas letras como o *j* [3], por exemplo, terem transcrições muito distintas em várias línguas conduziu à adição de novos símbolos fonéticos por parte de vários parceiros.

A colocação de marcas de silabificação em nomes de origem estrangeira levantou problemas muito interessantes, embora não fosse obrigatória no léxico inter-língua. De facto, os critérios de silabificação adoptados por cada parceiro para a sua própria língua diferiram bastante. Daí a dificuldade de nativizar por exemplo um nome de uma língua onde o critério é predominantemente marcado pela estrutura morfológica numa outra com um critério muito distinto. Independentemente do critério adoptado para a língua nativa, podem também ocorrer erros de silabificação por desconhecimento da morfologia da língua estrangeira.

5.3 Contexto

O grau de nativização de um nome estrangeiro depende bastante da situação em que é pronunciado e da pessoa com quem se está a falar. De facto, quando, por exemplo, se está a falar com uma pessoa que conhece mal a língua estrangeira, há uma grande tendência para uma forte nativização, mesmo em casos de boa competência de pronúncia. Como exemplo de um possível contexto em que se poderia produzir a pronúncia nativizada por defeito, tomemos a seguinte situação: "Imagine que leu sobre um dado lugar num guia turístico, lugar esse sobre o qual nunca tinha ouvido falar. Conhece o país, mas desconhece a língua. Liga para a sua agência de viagens local e diz: *Gostaria de ir a ....* Qual seria a sua pronúncia?".

6 CONCLUSÕES E PERSPECTIVAS FUTURAS

A contribuição mais significativa do projecto ONOMASTICA foi a criação de dicionários de pronúncia de nomes próprios em 11 línguas, para posterior divulgação em institutos de investiga-

ção e aplicação em múltiplos serviços por parte de indústrias e operadoras de telecomunicações.

Desenvolveram-se e compararam-se vários métodos de conversão automática grafema-fone, em particular os baseados em auto-aprendizagem e estudaram-se os problemas particulares levantados pela pronúncia de acrónimos e de nomes estrangeiros. Neste último âmbito, construiu-se também um léxico de pronúncia inter-língua que inclui 1000 nomes turisticamnte importantes de cada um dos 11 países do consórcio, com pronúncias nativizadas cruzadas em cada uma das línguas. Este léxico permitiu o estudo dos factores que influenciam a nativização e a comparação de diferentes graus de adaptação à estrutura sonora de línguas estrangeiras.

Apesar do projecto ter formalmente terminado em Junho passado, o trabalho prossegue pelo menos até 1997 num novo projecto financiado pelo Programa Europeu Copernicus, com a colaboração de novos parceiros que abordam problemas semelhantes nalgumas línguas da Europa Central e de Leste - Checo, Polaco, Romeno, Eslovaco, Servo-Croata, Ucraniano, Latviano, etc.

AGRADECIMENTOS

O projecto ONOMASTICA foi um esforço conjunto que envolveu muitos investigadores de várias instituições universitárias a quem os autores gostariam de agradecer. Este artigo menciona assim contribuições dos seguintes investigadores (por ordem alfabética): Ove Anderson (CPK), Lou Boves (Univ. Nijmegen), Paul Dalsgaard (CPK), Vassilis Darsinos (Univ. Patras), Bjorn Granstrom (KTH), Joakim Gustafson (KTH), Henk van den Heuvel (Univ. Nijmegen), Mervyn Jack (CCIR), George Kokkinakis (Univ. Patras), Emmy Konst (Univ. Nijmegen), Michael Logothetis (Univ. Patras), Andreas Mengel (Institut für Fernmeldetechnik), Peter Molbaek (CPK), Georg Ottensen (Sintef Delab), Jose Pardo (UPM), Vito Pirrelli (ICL), Mark Schmidt (CCIR), Andrew Sutherland (CCIR), Francisco Valverde (UPM), e François Yvon (Telecom Paris). Gostaríamos, em particular, de agradecer aos nossos colegas do INESC e do CLUL: Fernando Silva, Ernesto de Andrade, Ermelinda Gonçalinho e Catarina Moraes

REFERENCES

[1] O. Andersen e P. Dalsgaard, "A Self-Learning Approach to Transcription of Danish Proper Names", Proc. ICSLP'94, Yokohama, 1994.

[2] J. Esling, "Computer coding of the IPA: Supplementary Report", Journal of the International Phonetic Association, 20:1, 1990.

[3] J. Gustafson, "Transcribing names with foreign origin in the Onomastica Project", Proc. Int. Congress on Phonetics, Estocolmo, 1995.

[4] H. van den Heuvel, "Pronunciation of foreign names by Dutch grapheme-to-phoneme conversion rules", Proc. of the 2nd Onomastica Research Colloquium, Londres, 1994.

[5] A. Mengel, "Transcribing names - a multiple choice task: mistakes, pitfals and escape routes", Proc. of the 1st Onomastica Research Colloquium, Londres, 1993.

[6] V. Pirrelli e S. Federici, "You'd better say nothing than say something wrong: analogy, accuracy and text-to-speech applications", Proc. of the European Conf. on Speech Technology, Madrid, 1995.

[7] M. Plénat, "Observations sur le mot minimal en Français". In Laks & Plénat (eds), De Natura Sonorum, Presses Universitaires de Vincennes.

[8] M. Schmidt, S. Fitt, C. Scott e M. Jack, "Phonetic transcription standards for European names (ONOMASTICA)", Proc. of the European Conf. on Speech Technology, Berlim, 1993.

[9] T. Sejnowski e T. Rosenberg, "Parallel networks that learn to pronounce English text", Complex Systems, 1987.

[10] I. Trancoso, M. C. Viana, F. Silva, G. Marques e L. Oliveira "Rule-based vs. neural network-based approaches to letter-to-phone conversion for Portuguese Common and Proper Names", Proc. ICSLP'94, Yokohama, 1994.

[11] I. Trancoso (em nome do Consórcio ONOMASTICA), "The ONOMASTICA interlanguage pronunciation lexicon", Proc. of the European Conf. on Speech Technology, Madrid, 1995.

[12] M. C. Viana, I. Trancoso, F. Silva, G. Marques, E. Andrade e L. Oliveira, "Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu", Actas do Congresso Internacional sobre o Português, 1994.

[13] A. Villalva, "Compounding in Portuguese", Rivista di Linguistica, Vol. 4, no. 1, 1992.

[14] F. Yvon, "Régles de Transcription Graphème-Phonème pour la prononciation automatique des sigles", Lynx, 30 (no prelo).