

M. Céu VIANA
CLUL
Ernesto d'ANDRADE
FLL / CLUL
Luís C. de OLIVEIRA
INESC / IST
Isabel M. TRANCOSO
INESC / IST

Ler_PE : Um utensilio para o estudo da ortografia do Português

1. Introdução

Ler_PE é um sistema de transcrição fonética automática do Português Europeu, resultado do trabalho de colaboração entre o Grupo de Fonética e Fonologia do CLUL e o Grupo de Processamento de Fala do INESC, no âmbito do Projecto DIXI¹.

Como é bem conhecido, a arquitectura de um sistema de síntese de fala a partir do texto escrito tem necessariamente de obedecer a um certo número de requisitos de eficiência que, embora necessários para o seu funcionamento em tempo-real, nem sempre são compatíveis com a sua utilização directa como instrumento de investigação. Se essa arquitectura for suficientemente modular, o desenvolvimento de aplicações para fins de investigação aparece como um processo tanto mais natural quanto permite aproveitar alguns procedimentos que são necessários durante as fases de implementação e teste do sistema mas que têm de ser posteriormente retirados.

Desenvolvido a partir de conjuntos de regras de rescrita, que são basicamente as mesmas já utilizadas em Corso I², Ler_PE corresponde, na generalidade, ao módulo de transcrição fonética de DIXI e, tal como este, apresenta uma série de melhoramentos relativamente a essa primeira versão. Trata-se, no entanto, de um produto independente do sistema de síntese que tem por objectivo contribuir para um estudo estatístico da ortografia portuguesa e que pode ser utilizado na construção de dicionários fonéticos, ou acoplado a dicionários electrónicos para

gerar automaticamente as formas fonéticas de superfície, quer das formas de citação neles incluídas, quer, eventualmente, do texto que se lhes segue.

A arquitectura geral de Ler_PE está representada na figura 1. Procuraremos descrevê-la e fundamentá-la. Referiremos, também, os principais resultados obtidos durante o teste do sistema que sugerem algumas observações sobre a natureza da ortografia do Português.

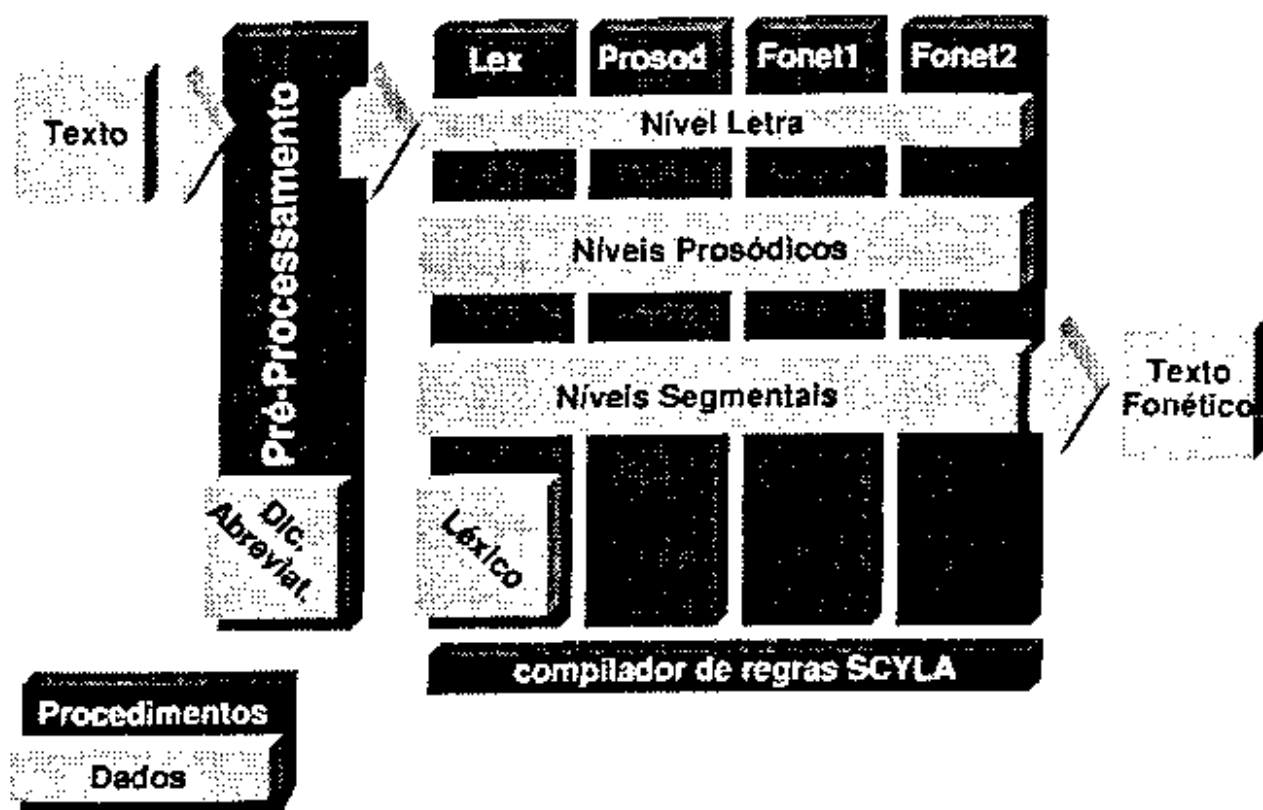


Figura 1: Diagrama de blocos do sistema

2. Descrição geral do sistema

O núcleo do sistema é constituído pelo conjunto de procedimentos que asseguram a transcrição fonética propriamente dita. Todos os outros conjuntos de procedimentos são de execução opcional.

No seu modo de funcionamento por defeito, todos os conjuntos são executados. O sistema permite visualizar as regras, analisar passo a passo os resultados da

aplicação de cada uma delas e, se necessário, alterá-los interactivamente. Quando este modo é desactivado, a aplicação dos diferentes procedimentos deixa de ser transparente para o utilizador, que fornece um texto ortográfico de entrada e obtém um texto fonético de saída.

Ler_PE admite, também, um modo de funcionamento estatístico, especialmente concebido para a análise de "corpora" de frequência. Neste caso, o texto ortográfico de entrada pode ser constituído por duas colunas de dimensão variável, a da direita contendo a(s) forma(s) a transcrever e a da esquerda a indicação da sua frequência de ocorrência ³.

2.1. Normalização do texto escrito

Nem todos os textos que temos de ler estão escritos em ortografia normalizada. Muitos contêm dígitos, abreviaturas e siglas, que é necessário desenvolver. Essa tarefa de tradução de todas as sequências de caracteres gráficos que não constituam palavras nem sinais de pontuação tem de ter em consideração o contexto em que essas sequências ocorrem. Assim, por exemplo, o símbolo "/" pode ser transcrito como "a dividir por" quando aparece numa expressão numérica (ex. $4/2=2$) ou interpretado como parte integrante de uma data (ex. 6/10/91).

Por razões de eficiência, este módulo é programado directamente em linguagem C, utilizando funções para a compilação e construção de expressões regulares ⁴.

2.1.1. Representação interna dos caracteres gráficos

Como a normalização de um texto envolve uma série de testes à estrutura segmental das diferentes formas que o constituem, o primeiro passo deste nível, que podemos considerar de pré-processamento da informação, consiste na conversão das maiúsculas em minúsculas e dos caracteres de oito bits utilizados pelo código ASCII estendido, numa representação interna de 7 bits. Esta característica é fundamental para qualquer sistema que tenha por objectivo correr em diferentes tipos de plataformas e pretenda tratar a ortografia de línguas que, tal

como a do Português, fazem apelo a cedilhas e a marcas gráficas de acento, normalmente codificadas com ASCII estendido. É que, embora exista uma norma internacional para este conjunto alargado de caracteres, esta nem sempre é respeitada pelos fabricantes de computadores ou de programas. Este facto dá origem a uma série de problemas sempre que se pretende transferir dados de um sistema para outro ou que os programas de introdução e processamento dos dados não utilizam a mesma representação interna. O módulo de normalização do texto escrito aceita assim, também, as marcas gráficas de acento e de cedilha separadas dos caracteres gráficos a que estão normalmente associadas, considerando como equivalentes representações tais como "á", "a" ou "a" e mudando, sempre que necessário, a posição destas marcas, de acordo com o seu formato de representação interna. Esta característica permite o tratamento de formas ou de textos introduzidos em computador com teclados estrangeiros ou com pacotes de programas que não suportam a totalidade dos caracteres portugueses.

2.1.2. Desenvolvimento de siglas, abreviaturas e dígitos

Uma normalização exaustiva e minimamente adequada do texto escrito é uma tarefa difícil que requer o recurso a informações semânticas e pragmáticas relativas ao contexto global de transmissão da informação e que obriga a certas decisões baseadas em critérios de ordem estatística, que merecem certamente um estudo mais aturado, mas que nos parece transcender, pelo menos de momento, os objectivos para os quais esta aplicação foi concebida.

O que nos parece fundamental assegurar num sistema como Ler_PE é uma arquitectura o mais modular possível e capaz de integrar com facilidade as informações que se julguem necessárias para o tratamento adequado do conjunto de dados em estudo.

O sistema contém assim, por exemplo, um pequeno dicionário de 95 expansões de abreviaturas que é consultado sempre que uma forma a ser analisada apresenta determinado tipo de terminação. Não se trata de um dicionário exaustivo das

expansões de abreviaturas possíveis em Português, mas apenas de uma amostra dessas expansões que poderá ser facilmente expandida ou modificada.

Ler_PE apresenta, no entanto, um certo número de limitações que são uma consequência directa de algumas decisões de ordem estatística, já tomadas. Por exemplo, para reduzir o risco de uma má tradução de expressões numéricas, apenas as datas válidas em formato numérico são traduzidas pelo sistema.

Foram também critérios de ordem estatística que serviram de base a uma tomada de decisão no que respeita ao tratamento de siglas e acrónimos. Os procedimentos de normalização soletram os que são constituídos apenas por sequências de consoantes e todos os outros são processados normalmente pelas regras gerais de transcrição fonética.

2.2. O léxico

Depois dos procedimentos de normalização, o sistema verifica para cada palavra da cadeia segmental a transcrever se esta faz, ou não, parte do léxico. Se tal se verificar, é-lhe associada a entrada lexical correspondente. Caso contrário, são aplicadas as regras de atribuição do acento lexical.

O léxico é constituído por um conjunto de formas que, de alguma maneira, implicam um tratamento particular. Algumas destas formas são verdadeiras excepções às regras de atribuição do acento, mas outras não. Estão no léxico, por exemplo, as homógrafas heterofónicas mais frequentes, assim como o conjunto de palavras gramaticais. Cada entrada no léxico tem a forma exemplificada em (1).

(1) {"quatro", 2, "_k_w_a_t_r_u, _qnt, _foc }

estando, em primeiro lugar, a forma ortográfica normalizada e, depois, o índice de posição do acento, a transcrição fonética e a indicação da categoria e do possível papel gramatical.

2.3. A atribuição do acento lexical

Por razões de eficiência, o conjunto de regras de atribuição do acento lexical foi escrito directamente em linguagem C, utilizando expressões regulares. Embora esta atribuição seja fundamental para uma correcta transcrição de todas as formas que não estejam contidas no léxico ou que não possuam marcas gráficas de acento, o conjunto de procedimentos que a realizam também é de execução opcional. Esta característica permite que o teste das regras do acento seja realizado de modo totalmente independente do das regras de transcrição fonética. Para tal, é apenas necessário que a marca de acento (") seja introduzida directamente pelo utilizador na cadeia segmental a transcrever, imediatamente antes da vogal acentuada:

(2) qu"atro

A possibilidade de desactivar este conjunto de procedimentos reveste-se ainda de algum interesse se o sistema for utilizado para fins de investigação, permitindo, por exemplo, o desenvolvimento e teste de outras regras de acentuação definidas pelo utilizador.

2.4. Transcrição fonética automática

Ao pretender transcrever foneticamente um texto escrito, somos obrigatoriamente confrontados com um certo número de opções que vão desde a escolha do alfabeto fonético ou do estilo a adoptar, até à do processo de transcrição mais adequado em face da natureza da ortografia da língua em estudo.

2.4.1. O alfabeto fonético

O Alfabeto Fonético Internacional (AFI) é, sem dúvida, o sistema de notação mais utilizado para transcrever foneticamente as mais variadas línguas do mundo. Para descrever as variantes fonéticas possíveis nessas línguas e permitir uma comparação entre elas, é necessário utilizar um grande número de símbolos e de diacríticos que não são directamente suportados pelos sistemas informáticos⁵. Os recentes desenvolvimentos no âmbito da engenharia da linguagem conduziram, entretanto, à definição de um conjunto de normas de notação fonética,

	AFI	SAM-PA	Exemplos	Transcrição
acento	[']	˘	pá	[p˘a]
nasalidade	[-]	˜	lã	[l˘6-]
vogais	[i]	i	vir, vinda	[v˘ir], [v˘i-d6]
	[e]	e	pera, pente	[p˘er6], [p˘e-t@]
	[ε]	E	leve	[l˘Ev@]
	[a]	a	cara	[k˘ar6]
	[æ]	6	câmara, canto	[k˘6m6r6]
	[i̯]	@	pequenina	[p@k@n˘in6]
	[ɔ]	O	corda	[k˘Ord6]
	[o]	o	dor, dom	[d˘or], [d˘o-]
	[u]	u	cume, mundo	[k˘um@], [m˘u-du]
semi-vogais	[j]	j	boi	[b˘oj]
	[w]	w	pau	[p˘aw]
consoantes	[p]	p	pai	[p˘aj]
	[t]	t	tia	[t˘i6]
	[k]	k	casa	[k˘az6]
	[b]	b	bote	[b˘Ot@]
	[d]	d	dote	[d˘Ot@]
	[g]	g	gato	[g˘atu]
	[β]	B	baba	[b˘ab6] ou [b˘aB6]
	[ð]	D	dada	[d˘ad6] ou [d˘aD6]
	[γ]	G	gaga	[g˘ag6] ou [g˘aG6]
	[f]	f	fava	[f˘av6]
	[s]	s	saca	[s˘ak6]
	[ʃ]	S	chave	[S˘av@]
	[v]	v	vaca	[v˘ak6]
	[z]	z	azul	[6z˘ul-]
	[ʒ]	Z	agir	[6Z˘ir]
	[l]	l	lado	[l˘adu]
	[l̥]	l̥	sal	[s˘al-]
	[λ]	L	folha	[f˘oL6]
	[r]	r	maré	[m6r˘E]
	[R]	R	carro	[k˘aRu]
[m]	m	mar	[m˘ar]	
[n]	n	neta	[n˘Et6]	
[ŋ]	J	spanha	[6p˘6J6]	

Quadro 1 - Lista de símbolos do SAM_PA utilizados na transcrição fonética do Português, correspondência com os símbolos do AFI e exemplos da sua utilização.

especialmente concebidas para o tratamento multilingue das línguas europeias por computador.

Ler_PE utiliza este novo alfabeto, o SAM-PA⁶, em vez do sistema de notação definido para Corso I que se veio a revelar desnecessariamente complexo e de muito difícil legibilidade. O quadro 1 apresenta a lista de símbolos do SAM-PA com que é realizada a transcrição fonética do Português.

2.4.2. O dialecto e o estilo

No desenvolvimento e teste do sistema utilizámos o recurso à nossa própria competência de falantes do Português para estabelecer as correspondências entre as representações ortográficas e as representações fonéticas, assim como as regras fonológicas propriamente ditas. A transcrição fonética realizada pelo sistema corresponde, pois, ao que podemos designar, em termos gerais, como uma variedade possível da região centro.

Em qualquer variedade são admissíveis, no entanto, diferentes estilos. Tendo em consideração os objectivos de Ler_PE acima enunciados, procurámos que a transcrição fonética correspondesse a uma leitura neutra, isto é, uma leitura não-casual, nem especificamente orientada para o ouvinte, relevando de um estilo formal. Parece-nos ser este o modo de transcrição mais adequado para o tratamento fonético de dicionários e que permite, simultaneamente, resultados mais aceitáveis na transcrição de textos livres.

Trata-se, na realidade, da opção por defeito de Ler_PE, que não é limitativa da capacidade do sistema como utensílio para a investigação. De facto, este foi concebido de modo a permitir a introdução de condições de aplicabilidade das regras aos vários níveis do processamento linguístico, sendo justamente exemplos dessas condições a variante dialectal e o estilo. Esta capacidade permitir-lhe-á vir a gerar pronúncias alternativas, desde que as regras específicas de outras variantes ou estilos sejam introduzidas.

2.4.3. O compilador de regras

Uma outra vantagem de Ler_PE relativamente a versões anteriores, como Corso I, é, sem dúvida, a utilização de SCYLA (de "Speech Compiler (for) Your Language"), um compilador de regras desenvolvido pelo CSELT⁷. Trata-se, mais precisamente, de uma linguagem de alto nível, adequada a uma análise multi-linear das estruturas de dados e cuja notação é muito semelhante à utilizada pela análise linguística, o que permite, por um lado, uma melhor estruturação e manipulação dos conjuntos de regras e, por outro, favorece o desenvolvimento do sistema por parte de linguistas não familiarizados com linguagens de programação.

Uma vez que a definição e sincronização dos diferentes níveis de análise são totalmente dependentes do utilizador, SCYLA facilita uma grande modularidade e versatilidade do sistema. Tem ainda a vantagem de gerar programas-fonte em linguagem C o que garante, por um lado, uma maior eficiência do sistema e, por outro, a sua portabilidade para qualquer processador que suporte esta linguagem.

Um programa em SCYLA consiste num bloco de declarações e num conjunto de procedimentos. No bloco de declarações são definidos os níveis escolhidos pelo utilizador, os objectos que pertencem a cada um destes níveis e os seus atributos. As regras que constituem os diferentes procedimentos são regras de rescrita com a forma:

$$A \rightarrow B / C \text{---} D$$

que se podem aplicar individualmente ou por conjuntos, linear ou ciclicamente. Cada elemento de uma regra pode corresponder apenas a um objecto do nível de análise ou a conjuntos de objectos que partilham um ou mais atributos. Pode ainda corresponder a uma estrutura lógica complexa e fazer, ou não, apelo a objectos ou estruturas de objectos de qualquer dos níveis de análise que foram definidos.

Como as regras de SCYLA podem ser combinadas com funções em linguagem C definidas pelo utilizador, a cada regra está associado um contador da frequência da

sua aplicação. Este encontra-se inactivo por defeito, sendo activado apenas quando o sistema corre em modo estatístico.

À semelhança do que acontecia já com a versão Corso I, o tratamento efectuado pela presente versão de Ler_PE obedece fundamentalmente (embora não exclusivamente) a critérios de ordem fonológica. O sistema parte do pressuposto de que existe uma correspondência directa entre a representação ortográfica e a representação fonológica, assumindo que a cadeia de elementos ortográficos de entrada constitui a sequência segmental básica de um sistema de representação multi-linear.

3. Principais resultados e dificuldades

O sistema foi testado com base num "corpus" de frequência constituído a partir do conjunto de entrevistas coligidas pelo CLUL para o Português Fundamental⁸. Este "corpus" contém cerca de 26 664 formas diferentes que correspondem a 715 686 ocorrências de formas de citação e de formas flexionadas e a 3 029 167 ocorrências de símbolos ortográficos. Durante a fase de teste foi utilizada uma versão do léxico com cerca de 300 entradas, apenas contendo as informações necessárias para o processamento prosódico.

3.1. Procedimentos de atribuição do acento

Na presente versão, o acento é correctamente atribuído em 99,9% dos casos, por um conjunto de 18 regras e os erros detectados dizem exclusivamente respeito a formas derivadas.

De facto, a posição do acento de palavra não pode ser determinada a partir de um conjunto de regras de base exclusivamente fonológica. Certos sufixos, bastante produtivos em Português, têm um comportamento especial, não implicando a desacentuação das palavras a que se associam. São exemplos destes sufixos - *zinho*, -*zito*, -*zão* e -*mente*. Para transcrever correctamente o conjunto de palavras derivadas com este tipo de sufixos é necessário atribuir-lhes duas marcas de

tonicidade. Este facto foi até 1973⁹ contemplado pela ortografia portuguesa que indicava com acento grave ou circunflexo a localização do acento na base, sempre que esta localização fazia apelo ao conhecimento da sua estrutura interna.

(3) (a) juridicamente (até 1973)

(b) juridicamente (depois de 1973)

Tendo estas marcas sido suprimidas, para acentuar correctamente a forma em 3(b), por exemplo, é necessário verificar, para além da terminação, a presença na base de outros sufixos que, tal como "ico", se caracterizam por um comportamento particular em relação às regras gerais de acentuação.

O sufixo "ico", que ocorre em formas como (3) não é, contudo, o mesmo que ocorre nas formas em (4) e em (5):

(4) burrico

mafarrico

(5) burricozinho

mafarricozinho

Para resolver casos de ambiguidade como estes, é utilizado um critério estatístico. Se for detectada a presença do sufixo "ico" na base, para este tipo de palavras derivadas, é pouco provável que este corresponda a um diminutivo, uma vez que as sequências de sufixos avaliativos, embora possíveis, não são frequentes.

O sistema ignorará a primeira vogal do sufixo, acentuando correctamente as formas em (3) e (4) e erradamente as formas em (5):

(6) burr^íco

mafarr^íco

(7) *b^íurricoz^íinho

*maf^íarricoz^íinho

Devido também à supressão das marcas de tonicidade da base em formas derivadas, são também erradamente acentuadas formas como :

(8) cafezinho - *c"afez"inho

alvarazinho - *alv"araz"inho

Se a frequência de ocorrência das formas erradamente acentuadas o justificar, estas poderão, contudo, ser incluídas no léxico pelo utilizador.

O interesse da utilização do sistema em modo estatístico não reside, no entanto, apenas na análise dos erros. Uma vez que a cada regra está associado um contador da sua frequência de aplicação, poderemos analisar, por exemplo, o peso relativo dos diferentes tipos de conhecimento linguístico a que é feito apelo

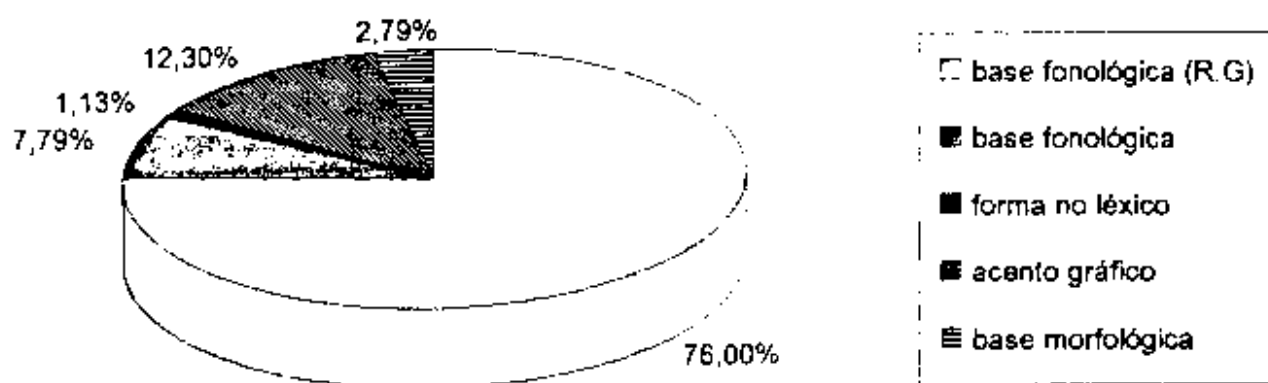


Figura 2: Frequência relativa dos diferentes tipos de regras de acentuação

A figura 2 ilustra o peso relativo de quatro grandes grupos de regras: (1) as que se limitam a interpretar marcas gráficas de acento, (2) as que associam uma dada forma a uma entrada lexical, (3) as que fazem apelo a conhecimentos de ordem morfológica (4) as que se baseiam exclusivamente em critérios de ordem fonológica. Como a mesma figura mostra, entre as regras de base fonológica, a regra geral de acentuação, que atribui o acento à segunda vogal a contar do fim, aplica-se, só por si, a 76% das ocorrências. A percentagem global de erros na atribuição do acento é,

como já referimos, diminuta e diz respeito unicamente a uma análise morfológica incorrecta.

3.2. Procedimentos de transcrição fonética

O primeiro passo do sistema de regras consiste no preenchimento do primeiro nível de análise com a informação contida no léxico e com as marcas de acento atribuídas pelas regras de acentuação. A transcrição fonética propriamente dita é realizada por dois conjuntos de procedimentos independentes mas que operam sobre os mesmos níveis de estruturas de dados. A cadeia segmental é percorrida da esquerda para a direita, sendo a transcrição fonética realizada palavra a palavra, ao nível da forma de citação. A sequência em (9) será primeiro transcrita como (10) e só numa segunda etapa é que serão tratados os encontros vocálicos e consonantais em fronteira de palavra, obtendo-se transcrições como a de (11).

(9) Ele disse à Antónia que vinha amanhã

(10) *el@ d*is@ a 6~t*Onj6 k@ v*inh6 am6J*6~

(11) el~d*is a~t*Onj6 k v iJ am6J6~

Este modo de operar, embora não seja imprescindível para a produção de resultados correctos, permite uma maior clareza e simplicidade na ordenação das regras e na verificação das suas condições de aplicabilidade. Permite, por outro lado, que sejam geradas transcrições fonéticas ao nível da soletração em que são preservadas todas as vogais átonas. Estas são conservadas ou suprimidas dependendo do estilo seleccionado e da estrutura prosódica do enunciado, podendo algumas regras que se aplicam a este nível desencadear processos de ressilabação.

Para o "corpus" de teste, estes conjuntos de regras produzem 98% de resultados fonéticos correctos, ao nível da transcrição fonética dos segmentos e 91,6% ao nível da palavra. Estes resultados, assim como a análise dos erros, vêm corroborar algumas observações anteriores baseadas na análise do vocabulário do Português Fundamental, contendo apenas 2217 formas de citação ¹⁰. De facto, a maior parte

dos erros corresponde a uma má transcrição de "e" e "o" quando tónicos e devem-se, sobretudo, à dificuldade de realizar uma análise morfológica das formas contidas no "corpus", sem consulta a um dicionário de morfemas. Esta análise é necessária uma vez que a ortografia portuguesa não tem em consideração o facto de fonologicamente /ε / se opôr a /e/ e /ɔ/ a /o/ mas não resolve todos os casos de ambiguidade fonológica, nomeadamente a que se verifica para as homógrafas heterofónicas, que constituem cerca de 3% do "corpus" de teste.

Tal como para as regras de acentuação, também aqui podemos analisar o peso relativo dos diferentes tipos de conhecimento linguístico a que a transcrição fonética automática do texto escrito tem de fazer apelo. A figura 3 ilustra a frequência de aplicação para o "corpus" de teste de diferentes tipos de regras: (1) as que determinam a realização fonética dos segmentos independentemente do contexto, (2) as que são dependentes do contexto e de base estritamente fonológica, (3) as que fazem apelo a conhecimentos de ordem morfológica e (4) as que têm em consideração, para além das propriedades fonéticas a nível segmental, a estrutura prosódica do enunciado.

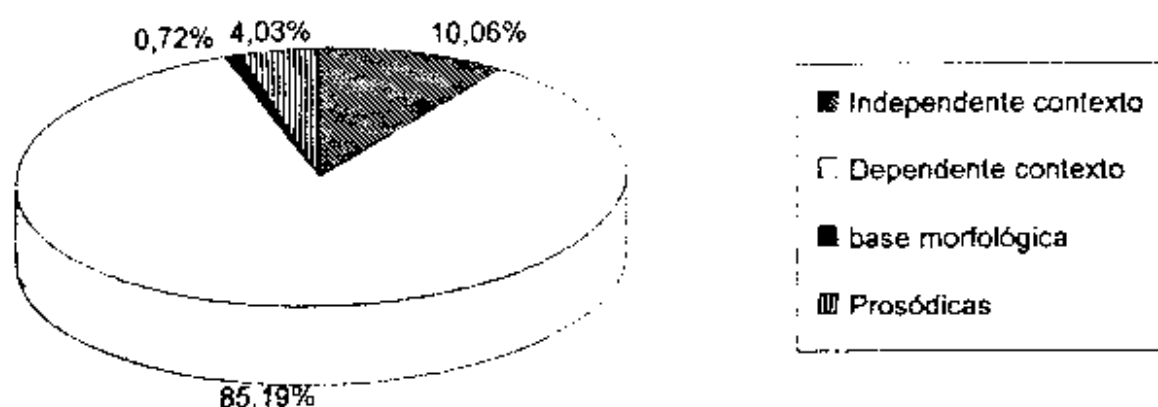


Figura 3: Frequência relativa dos diferentes tipos de regras de transcrição fonética.

4. Considerações finais

Apesar destas dificuldades, os resultados obtidos sem recurso a listas de excepções mostram que a ortografia portuguesa pode ser considerada de base fundamentalmente fonológica e justificam o tratamento adoptado. A maior parte das regras de acentuação ou de transcrição que fazem apelo a conhecimentos de ordem morfológica são baseadas em critérios de ordem estatística e são totalmente desnecessárias para algumas aplicações para as quais o sistema foi concebido. Elas podem ser suprimidas, por exemplo, se o sistema vier a ser acoplado a um dicionário electrónico, pois este poderá fornecer a informação necessária sobre a estrutura interna da palavra ou sobre qual a vogal temática do verbo, reduzindo substancialmente o número de erros cometidos pelo sistema.

Um aspecto que nos parece interessante referir é que um sistema como Ler_PE pode ser encarado como uma base de conhecimentos sobre as relações entre a ortografia e o som, pelo que os resultados de uma análise estatística de um "corpus" de frequência como o do Português Fundamental podem também ser considerados de um ponto de vista da pedagogia da leitura do Português como primeira ou segunda língua.

Como tivemos ocasião de referir, a maior parte dos conhecimentos representados em Ler_PE são de base fonológica e dizem respeito a uma só variedade do Português. O sistema encontra-se preparado, contudo, para incorporar outras variedades desta língua e outros estilos mais ou menos formais. É nossa intenção explorar essas capacidades do sistema, aumentando a sua gama de conhecimentos e utilizando-o como instrumento de investigação sobre as semelhanças e as diferenças entre estilos e variantes do Português.

Bibliografia

- Andrade, E. e M. Céu Viana (1985) - Curso I - Um conversor de texto ortográfico em código fonético para o Português. *RGFF*, 5. CLUL-INIC.
- Carvalho, P., P. Geada e P. Lopes (1991) - "Norm: Normalizador de Texto para Português". *Relatório interno do INESC*.
- Lazzaretto, S. e L. Nebbia (1987). "SCYLA: Speech Compiler for Your Language", *Proceedings of the 1st European Conference on Speech Communication and Technology*, vol 2: 381-384.
- Nascimento, F., L. Marques e L. Segura (1987) - *Português Fundamental: Métodos e Documentos*. INIC-CLUL, Lisboa.
- Oliveira, Luis C., M. Céu Viana e Isabel Trancoso (1991) - "DIXI - Portuguese Text-to-Speech System". *Proceedings of the 2nd European Conference on Speech Communication and Technology*, vol. 3: 1239-1242.
- Winsky, R., W. J. Barry e A. Fourcin (eds) (1989) - Support Available from SAM Project for other Esprit Speech and Language Work, Esprit Project 2589 (SAM), Multi-Lingual Speech Input/Output Assessment, Methodology and Standardisation.

Notas

¹ DIXI é um sistema de síntese do Português a partir do texto escrito concebido como uma estrutura suficientemente modular e flexível, capaz de ser utilizada no seu todo, ou apenas em parte, como um utensílio para a investigação em linguística e fonética e de permitir o desenvolvimento e a avaliação de novos modelos de produção de som. Para a descrição detalhada dos objectivos e da arquitectura geral do sistema cf. Oliveira, Viana e Trancoso (1991).

² Cf. Andrade e Viana (1985)

³ Note-se que o sistema pode operar em modo estatístico sobre qualquer tipo de texto em formato ASCII. Neste modo, no entanto, qualquer sequência de dígitos em início de linha é interpretada como indicadora da frequência de ocorrência das formas contidas nessa mesma linha. Na ausência de uma sequência de dígitos nesta posição é assumido o valor por defeito (=1).

⁴ Cf. Carvalho, Geada e Lopes (1991)

⁵ Embora alguns editores de texto suportem já o conjunto de símbolos gráficos que constituem o Alfabeto Fonético Internacional, estes são geridos internamente e não são suportados pelo sistema operativo.

⁶ Cf. Winski, Barry e Fourcin (1989)

⁷ Cf. Lazzaretto e Nebia (1987)

⁸ Cf. Nascimento, Marques e Segura (1987).

⁹ Cf. Decreto-Lei nº 32/73

¹⁰ Cf. Andrade e Viana (1985)