Fraseamento Prosódico: uma Experiência de Avaliação

M. Céu Viana
CLUL
Luís C. Oliveira
INESC-ID/IST
A. Isabel Mata
FLUL/CLUL

1. Introdução

No desenvolvimento de um sistema de síntese a partir de texto, os métodos de avaliação têm um papel fundamental tanto na fase de optimização dos modelos como, posteriormente, para avaliar a capacidade de generalização destes.

O processo de avaliação mais comum, geralmente utilizado durante as fases de optimização dos modelos, é o que compara os valores preditos com valores de referência observados para um determinado *corpus*. Este processo, que tem a vantagem de ser integralmente automático, foi o utilizado no desenvolvimento do módulo de fraseamento prosódico a integrar na nova versão do sistema DIXI (DIXI+), desenvolvido no âmbito do acordo de cooperação entre o CLUL e o INESC (cf. Viana e Mata, 2001). Contudo, as medidas de desempenho calculadas deste modo podem ser enganadoras, uma vez que diferentes partições prosódicas dos enunciados, particularmente dos mais longos, podem ser igualmente aceitáveis para os falantes/ouvintes. Avaliações mais realistas são evidentemente possíveis quando estão disponíveis vastos conjuntos de materiais de fala, previamente anotados, contemplando múltiplas produções de cada enunciado por um mesmo falante e/ou por um número razoável de falantes. Não estando ainda reunidas estas condições para o Português Europeu (PE), tal como para muitas outras línguas, foi necessário encontrar estratégias de avaliação alternativas.

Para esse efeito, e de modo a angariar mais facilmente um número significativo de avaliadores, procurou-se desenvolver um teste que utilizasse apenas texto, não fosse excessivamente demorado e pudesse ser realizado à distância através da internet, com base numa ferramenta especialmente desenvolvida para esse efeito.

2. O teste de avaliação

Para além de verificar se as partições prosódicas preditas pelo modelo de fraseamento poderiam ou não corresponder a leituras possíveis de cada um dos enun-

Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2002, pp. 517-529

ciados, pretendia-se avaliar também a adequação das anotações utilizadas nos materiais de treino, uma vez que, ao contrário do que é habitual no desenvolvimento deste tipo de sistemas, estas não resultaram da audição e observação das propriedades presentes no sinal acústico mas foram marcadas directamente sobre o texto, adoptando uma estratégia proposta e testada com sucesso para o Inglês e o Espanhol por Hirschberg e Prieto (1996).

2.1. Objectivos

O desenvolvimento do teste teve em mente, por conseguinte, três objectivos distintos e complementares:

1. Avaliar a aceitabilidade das rupturas atribuídas pelo sistema automático;
2. Avaliar a opinião dos avaliadores sobre as rupturas usadas como referência para o treino do sistema automático;
3. Avaliar a variabilidade dos avaliadores na tarefa de segmentação prosódica dos enunciados.

2.2. Tarefas dos avaliadores

Uma vez que se pretendia utilizar apenas texto, os dois primeiros objectivos podiam ser atingidos de forma idêntica, apresentando os enunciados escritos com as respectivas marcas de ruptura e pedindo aos participantes para os classificarem em três categorias distintas:

BOM (B) – Eu poderia ler desta forma;
ACEITÁVEL (A) – Eu não leria desta forma mas é uma leitura possível;
INACEITÁVEL (I) – Não me parece corresponder a uma leitura natural da frase.

Quanto ao terceiro objectivo, bastaria apresentar o texto sem quaisquer marcas de ruptura e pedir aos participantes que introduzissem as que julgassem mais adequadas para dar conta de uma leitura pausada mas clara e fluente.

Uma vez que se optou por recorrer a informantes não-especialistas, não se poderia esperar que estes fossem capazes de estabelecer distinções claras entre diferentes níveis de ruptura prosódica. Ora, como o sistema foi treinado para atribuir rupturas maiores e menores (embora não especificando o seu grau) seria necessário que os avaliadores pudessem operar de modo equivalente, sem que essa distinção fosse explicitada nas instruções. Solicitar-lhes uma leitura pausada pareceu-nos ser a melhor estratégia, dado que neste tipo de leitura as fronteiras de constituintes entoacionais maiores e menores são geralmente percebidas como estando associadas a pausas, independentemente de estas serem efectivamente realizadas ou não. Uma experiência piloto realizada antes da execução do teste mostrou, no entanto, que esta estratégia era ainda insuficiente por se verificar uma grande variabilidade

na partição prosódica dos enunciados: enquanto alguns avaliadores assinalavam as fronteiras de constituintes entoacionais menores e maiores, outros apenas marcavam estas últimas. Para ultrapassar esta dificuldade e permitir uma melhor comparação entre as partições prosódicas realizadas pelos avaliadores, as utilizadas como referência e as preditas pelo modelo automático, foram impostos limites máximos e mínimos ao número de rupturas a atribuir a cada enunciado. Este deveria situar-se dentro de uma gama pré-definida de possibilidades, calculada para cada enunciado com base nas partições observadas no *corpus* de referência e nas atribuídas pelo sistema automático. Uma vez que essa solução seria demasiado restritiva, principalmente quando o número de partições é idêntico nos dois casos, optou-se por permitir em todos os enunciados a introdução de menos uma ruptura do que o mínimo anteriormente calculado.

2.3. Selecção e distribuição dos enunciados

Querendo manter a dimensão do teste final dentro de limites razoáveis para que esta não constituísse um factor de dissuasão, impunha-se estimar em primeiro lugar o tempo médio necessário para a execução das diferentes tarefas e, em função deste, decidir qual o número de enunciados a incluir em cada uma delas.

De acordo com os resultados dos testes preliminares realizados para esse efeito, a avaliação de cada enunciado demoraria cerca de 1 minuto em média, pelo que seria conveniente restringir o trabalho de cada avaliador a cerca de 30 enunciados. Sendo este número muito inferior ao que, de acordo com as nossas observações, seria necessário para assegurar um mínimo de heterogeneidade nos materiais do teste (cerca de 90), foi necessário recorrer a uma estratégia para a atribuição dos enunciados a cada avaliador.

Os 90 enunciados a avaliar foram extraídos aleatoriamente de um conjunto de materiais de teste, depois de se terem eliminado todos os que, por serem relativamente curtos, são muitas vezes produzidos como um único constituinte entoacional. Tendo sido retirados do conjunto de teste, não foram usados para o treino da árvore de classificação que assegura o fraseamento automático, um requisito que é fundamental quando se pretende testar a capacidade de generalização de um modelo. Os enunciados seleccionados para avaliação têm no mínimo 7 e no máximo 65 palavras, sendo o seu comprimento médio de 19 palavras. Incluindo a ruptura final, o fraseamento de referência atribuiu neste conjunto um mínimo de 2 rupturas a enunciados mais curtos e um máximo de 16 aos mais longos.

Tar0	Tar1	Tar2
conj0r	conj1r	conj2r
conj1a	conj2a	conj0a
conj2m	conj0m	conj1m

Tar3	Tar4	Tar5
conj3r	conj4r	conj5r
conj4a	conj5a	conj3a
conj5m	conj3m	conj4m

Tar6	Tar7	Tar8
conj6r	conj7r	conj8r
conj7a	conj8a	conj6a
conj8m	conj6m	conj7m

Quadro 1 – Distribuição dos 270 enunciados por 9 tarefas, sub-divididas em conjuntos de 10, com as rupturas atribuídas automaticamente (a) as de referência (r) e texto simples para a inserção de marcas de ruptura (m).

Estes 90 enunciados foram então divididos em conjuntos de 10 (conj0 a conj8) e para cada um deles foram produzidas três versões: uma com as partições produzidas pelo sistema automático (a), outra com as partições de referência (r) e ainda outra, sem partições, para ser marcada pelos avaliadores (m). Os 270 enunciados resultantes foram em seguida distribuídos por 9 tarefas de acordo com quadro1.

As trinta frases de cada tarefa foram ainda ordenadas aleatoriamente para misturar os enunciados a avaliar com os enunciados a segmentar, procurando impedir, assim, tanto a habituação a um tipo de tarefa como a identificação por parte do avaliador dos dois tipos de fraseamento (o automático e o de referência).

2.4. Ferramenta de avaliação e recrutamento dos avaliadores

A ferramenta desenvolvida para a realização do teste utilizou a "*Common Gateway Interface*" (CGI) para, através de um servidor de HTTP *(Hyper Text Transfer Protocol*), gerar os formulários a preencher pelo utilizador. Desta forma, bastava que o avaliador tivesse acesso à Internet através de um "*web browser*", como o Netscape ou o Internet Explorer, para que pudesse participar no teste.

Os avaliadores foram convidados a participar através de mensagens de correio electrónico que incluiam o endereço URL (*Uniform Resource Locator*) do teste e que foram enviadas para o CLUL, a FLUL, o INESC-ID, o IST e a uARTE- MCT, bem como para alguns contactos pessoais dos investigadores envolvidos. Procurou-se também utilizar o efeito de *bola de neve*, pedindo aos próprios avaliadores para divulgarem o endereço do teste.

2.5. Execução do teste

Cada avaliador preenchia em primeiro lugar um formulário onde lhe era pedido que se identicasse, indicando o nome pelo qual queria ser conhecido e, em

seguida, o seu nome completo[1]. Findo este processo, era-lhe atribuída pelo sistema uma das nove tarefas possíveis, e dito que a partir daquele momento poderia continuar ou interromper a execução do teste em qualquer altura e retomá-la mais tarde no ponto em que a tinha largado, desde que se identificasse correctamente.

As tarefas iam sendo atribuídas à medida que cada novo participante se indentificava, de modo a que, uma vez terminadas as que estavam a cargo dos primeiros 9 avaliadores, houvesse uma resposta para as três versões de cada um dos 90 enunciados, duas respostas uma vez terminadas as tarefas atribuídas aos primeiros 18, e assim sucessivamente.

Quanto ao teste propriamente dito, cada tarefa era composta, como já foi referido acima, por 30 enunciados distintos, cada um apresentado em sua página. Nas 20 páginas correspondentes à avaliação das rupturas de referência ou das rupturas atribuídas automaticamente, o formato era idêntico para que não fosse possível a identificação de umas e outras. O texto era apresentado com as respectivas rupturas já marcadas e os participantes podiam carregar em três botões diferentes para lhes atribuir o tipo de classificação que julgassem mais adequada (**B**(om), **A**(ceitável) e **I**(naceitável)). Nos 10 enunciados restantes, era colocado um botão em cada fronteira de palavra, bastanto carregar em qualquer deles para que aparecesse (ou desaparecesse) uma marca de ruptura nessa fronteira. Uma vez que foi decidido impor limites máximos e mínimos ao número de rupturas a atribuir a cada enunciado, não era permitido avançar no teste se essa condição não fosse preenchida.

Dos 105 participantes no teste, 91 realizaram integralmente a tarefa que lhes tinha sido atribuída, havendo por conseguinte pelo menos 10 julgamentos para cada um dos 180 enunciados cuja partição se pretendia avaliar e 10 respostas para cada um dos 90 enunciados a segmentar.

O teste decorreu entre 27 de Março e 3 de Abril de 2001. Após o dia 3 de Abril, o URL do teste foi substituído por um programa que permitia a cada avaliador visualizar as suas próprias respostas e compará-las com as dos restantes participantes na mesma tarefa. Pretendeu-se desta forma fornecer aos avaliadores algumas informações susceptíveis de os interessar e que pudessem funcionar como incentivo para a continuidade da sua colaboração em iniciativas semelhantes a realizar futuramente.

[1] O pedido do nome completo tinha apenas por objectivo estimar a proporção de linguistas e não-linguistas e facilitar futuros contactos para outro tipo de testes. Entre os 90 avaliadores que colaboraram no teste, apenas nos foi possível identificar um número muito reduzido de linguistas. Não nos é possível, no entanto, estimar qualquer tipo de proporção, uma vez que a grande maioria dos avaliadores utilizou nomes falsos (ex. xfgsh jkasern). Entre estes, encontrava-se, certamente, um número indeterminado de linguistas, uma vez que, na fase de divulgação dos resultados do teste, fomos contactados por alguns que se encontravam nesse caso e pretendiam ter acesso a esses dados.

3. Resultados ao nível do enunciado

A grande maioria dos testes de avaliação de sistemas automáticos de partição prosódica apresenta os resultados ao nível do número de rupturas correcta ou incorrectamente atribuídas. Embora os resultados obtidos deste modo para o PE sejam em tudo comparáveis aos de outras línguas, nomeadamente aos de Taylor e Black (1998) para o Inglês, é fundamental analisar o comportamento do sistema não apenas ao nível das rupturas mas também ao nível do enunciado. Esta último nível é muito mais exigente porque um erro na atribuição de uma única ruptura pode ser suficiente para que o enunciado seja rejeitado como um todo. Este efeito torna-se evidente se analisarmos os resultados obtidos a estes dois níveis para os materiais utilizados no teste: enquanto que ao primeiro nível o sistema atribui cerca de 80% de rupturas correctas e classifica acertadamente cerca de 91% das fronteiras, o fraseamento automático do enunciado só coincide com o de referência em apenas 20 dos 90 enunciados considerados. Ou seja, com base neste critério, o sistema asseguraria apenas 22% de resultados correctos.

Qualquer modelo pressupõe, no entanto, alguma capacidade de generalização sobre dados, pelo que uma não coincidência entre o fraseamento automático e o de referência não significa, necessariamente, que o primeiro esteja incorrecto: todos os enunciados, particularmente os mais longos, são passíveis de diferentes leituras igualmente aceitáveis de um ponto de vista do falante/ouvinte. Uma das vantagens do presente teste de avaliação é, justamente, a de fornecer alguns elementos importantes para uma análise dos resultados a este nível. Dispõe-se agora de pelo menos 10 exemplos de atribuição de rupturas a cada um dos enunciados do teste e sabe-se qual é a opinião de pelo menos 10 avaliadores sobre a qualidade dos dois tipos de fraseamento prosódico. A adopção de qualquer medida de desempenho implica, contudo, uma análise prévia do comportamento dos avaliadores.

3.1. Variabilidade entre os avaliadores

Não sendo o fraseamento prosódico independente de factores como a interpretação sintáctico-semântica efectuada pelos falantes/ouvintes e o seu próprio estilo de leitura, seria de esperar uma considerável variação nas segmentações e nos julgamentos efectuados pelos diferentes avaliadores. Para analisar o desempenho destes nas diferentes tarefas que lhes foram atribuídas, consideraram-se inicialmente apenas os 20 enunciados em que o fraseamento automático é igual ao de referência, para os quais se dispõe da opinião de 20 avaliadores. Os resultados obtidos para esse conjunto enunciados são os apresentados no quadro 2.

Nesses 20 enunciados, apenas um (o enunciado 543) foi unanimente considerado como bom, embora apenas 60% dos participantes o tenham segmentado do mesmo modo, assinalando os restantes sobretudo a ruptura maior coincidente com a vírgula.

543: "Quando se tira o bolo do forno, / rega-se com este xarope / e polvilha-se com acúcar fino."

Mesmo quando a grande maioria dos avaliadores classifica uma partição como boa, há sempre 1, pelo menos, que não pensa desse modo. É o caso, por exemplo, do enunciado 8, em que a segmentação proposta é igual à de 90% dos avaliadores, mas rejeitada por um deles, tendo outro preferido uma leitura apenas com uma partição entre o sujeito e o predicado.

8: "Os clichés sexuais / e as distinções sociais / não são muito relevantes."

Sobretudo nos enunciados mais curtos, a tendência para utilizar o menor número possível de rupturas é manifesta[2]. Como 418 mostra, mesmo com constituintes antepostos, onde se esperaria uma ruptura (assinalada até por uma vírgula no texto original), 20% dos avaliadores trata este enunciado como um único constituinte entoacional (cf. 418b).

418a: "Com um aceno de cabeça / mandou-a deitar-se na esteira."
418b: "Com um aceno de cabeça mandou-a deitar-se na esteira."

O enunciado 18, em que 1 dos avaliadores também considera preferível não fazer qualquer partição interna (cf. 18b), apresenta apenas uma ruptura entre o sujeito e o predicado que suscita alguma variabilidade nos julgamentos. Muito embora 90% dos avaliadores marquem essa mesma ruptura, 20% rejeitam-na e apenas 40% a consideram boa. Na totalidade do *corpus* de teste, há bastantes casos deste tipo, sujeitos a idêntica variabilidade nos julgamentos, em que, por um lado, o enunciado parece demasiado longo para ser realizado como um único contituinte entoacional mas, por outro, uma ruptura preferencialmente associável a uma determinada fronteira cria uma assimetria evidente entre a extensão dos constituintes à esquerda e à direita do verbo.

18a: Este tipo de emoção/ goza de grande popularidade junto dos jovens./
18b: Este tipo de emoção goza de grande popularidade junto dos jovens./

[2] Apesar the alguns avaliadores manifestarem quase sempre esta tendência, recebemos alguns comentários lamentando não ser possível dispor de um maior número de rupturas, sobretudo para os enunciados mais longos. Este trabalho não seria possível, contudo, se tivessemos optado por pemitir um maior grau de liberdade: o número de combinações cresceria exponencialmente, dificultando (ou impedindo mesmo) uma comparação efectiva dos diferentes tipos de dados.

enunciado	extensão	seg. aval.	classif. B	classif. A	classif. I
8	11	90%	85%	10%	5%
18	11	90%	40%	40%	20%
43	19	0%	55%	40%	5%
53	30	0%	30%	40%	30%
103	23	40%	40%	45%	15%
118	8	70%	80%	15%	5%
133	13	50%	85%	10%	5%
173	8	90%	85%	5%	10%
378	17	10%	40%	45%	15%
383	31	10%	30%	35%	35%
398	17	60%	55%	35%	10%
418	9	80%	95%	5%	0%
438	12	80%	90%	10%	0%
443	11	90%	55%	25%	20%
458	16	80%	80%	10%	10%
473	9	90%	80%	20%	0%
488	9	80%	75%	25%	0%
538	11	100%	80%	20%	0%
543	16	60%	100%	0%	0%
548	34	0%	20%	35%	45%

Quadro 2: 20 frases cujo fraseamento automático foi igual ao de referência e respectiva classificação pelos 20 avaliadores.

Por vezes, é o próprio texto que impõe essa assimetria, como em 443, em que a segmentação proposta se limita a respeitar a pontuação do próprio autor, tal como o fazem 90% dos avaliadores. Apesar disso, 20% consideram esta partição como inaceitável e apenas 55% como boa, tendo um único avaliador proposto uma segmentação diferente (cf. 443b) que não parece constituir uma melhor alternativa. O que parece estar em causa aqui não é propriamente a adequação do modelo de fraseamento mas sim a própria frase, tal como foi escrita pelo seu autor, ou porque a consideram inaceitável ou porque não encontram um padrão melódico que justifique essa leitura

443a: "O que decepcionou o espectador, / foi, / parece-nos,/ a tibieza do texto./"
443b: "O que decepcionou o espectador, foi, / parece-nos,/ a tibieza do texto./"

Podem encontrar-se, contudo, exemplos bastante mais complicados, sobretudo em enunciados longos, como 548, em que as partições prosódicas dos avaliadores (que, por falta de espaço, não reproduzimos integralmente) nunca coincidem com a de referência e apenas coincidem entre si em dois casos (cf 548b).

548a: "Quantos pais/ não tentam constantemente/ cumprir as já velhas teorias/ que defendem o diálogo,/ em substituição da sevícia/ ou do castigo,/ e se angustiam/ quando a mão lhes foge,/ pesada,/ em direcção aos filhos?/ "

548b: "Quantos pais não tentam/ constantemente/ cumprir as já velhas teorias/ que defendem o diálogo,/ em substituição da sevícia/ ou do castigo,/ e se angustiam/ quando a mão lhes foge,/ pesada,/ em direcção aos filhos?/ "

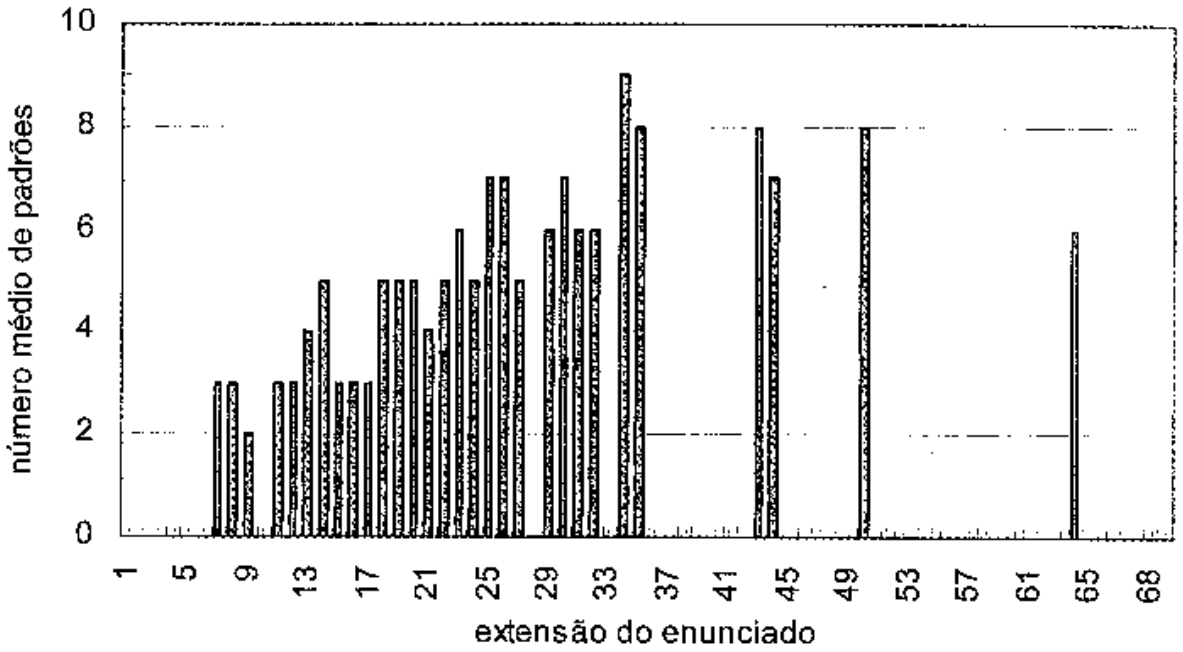

Figura 1: Número médio de padrões de segmentação diferentes assegurados por 10 avaliadores para enunciados com diferentes extensões no *corpus* de teste.

Este é talvez um exemplo extremo mas que ilustra a importância da extensão dos enunciados no grau de coerência inter-avaliadores. Como a figura 1 mostra, o número de partições possíveis de um mesmo enunciado tende a crescer à medida que a sua extensão aumenta e a partir de 29 palavras os 10 avaliadores produzem cerca de 6 ou mais padrões de fraseamento diferentes.

Embora, como já se mostrou acima, se possa observar uma importante variação nos julgamentos de alguns enunciados relativamente curtos e com um grau de concordância inter-avaliadores elevado, há sem dúvida uma relação clara entre ambos. Como o quadro 3 mostra, as taxas de rejeição tendem a aumentar quando o grau de concordância na segmentação feita pelos avaladores decresce. Esta regularidade é tanto mais importante quanto a avaliação e a partição prosódica de um mesmo enunciado nunca são asseguradas pelo mesmo grupo de avaliadores, permitindo pôr em evidência um forte grau de consistência entre eles que, de alguma forma, legitima a metodologia adoptada nesta experiência e contribui para uma melhor definição dos critérios de medida a adoptar.

De acordo com as observações realizadas, apenas parece razoável considerar um fraseamento *Inaceitável* se mais de metade dos avaliadores o tiver considerado como tal e *Bom* se assim tiver sido classificado também pelo mesmo número de avaliadores.

Partições idênticas	Avallação		
	Bom	Aceitável	Inaceitável
100%	80%	20%	0%
90%	71%	19%	10%
80%	80%	17%	3%
70%	67%	23%	10%
60%	73%	15%	12%
50%	68%	17%	15%
40%	54%	30%	16%
30%	38%	38%	23%
20%	43%	26%	31%
10%	36%	39%	25%
0%	20%	35%	45%

Quadro 3: Percentagem de avaliadores que realiza uma partição idêntica dos enunciados e respectivas avaliações dessas partições ao nível do enunciado

3.2. Resultados do fraseamento automático

Utilizando o critério que se acabou de definir, os avaliadores consideraram como inaceitáveis 20 dos 90 enunciados fraseados automaticamente (22%). Nestes enunciados incluem-se os dois mais longos do teste e um cujo fraseamento de referência foi considerado inaceitável. Dos restantes enunciados, 40 (44%) foram julgados aceitáveis e 30 (33%) bons. A título de referência pode indicar-se que em 40 (44%) enunciados houve pelo menos um avaliador que fez um fraseamento igual ao do sistema automático.

3.3. Resultados do fraseamento de referência

Aplicando os mesmos critérios para julgar a qualidade do fraseamento de referência utilizado para treino do sistema automático, verifica-se que o fraseamento de 6 (7%) enunciados foi considerado inaceitável, 31 (34%) bom e 53 (59%) aceitável. Neste caso, o número de enunciados em que um avaliador fez um fraseamento igual ao de referência subiu para 50 (56%).

4. Resultados ao nível da ruptura

Como Huang et. alii (2001) referem '*there are many reasonable places to pause in a long sentence, but few where it is critical not to pause*'. Os resultados que acabámos de apresentar põem em evidência, de facto, que a determinação do número de rupturas mal atribuídas não pode ser feita por simples comparação com o fraseamento de referência, uma vez que partições alternativas podem ser consideradas aceitáveis ou mesmo boas. É necessário determinar ainda, no entanto, se os enunciados rejeitados pela maioria dos avaliadores o são, simplesmente porque o sistema atribui rupturas em fronteiras onde não é possível fazê-lo ou também, como uma primeira análise dos dados sugere, porque não foi atribuída uma ruptura onde seria imprescindível introduzi-la.

Para procurar responder a estas questões, os dados foram analisados tendo em consideração três medidas de desempenho propostas em Taylor e Black (1998) e adoptadas em Viana e Mata (2001) para a optimização dos modelos, mas definidas agora de modo bastante diferente:

Ruptura correcta – sempre que pelo menos um dos avaliadores inseriu uma ruptura numa dada fronteira;
Falsa inserção – sempre que foi atribuída uma ruptura numa fronteira em que nenhum avaliador o fez;
Apagamento – sempre que não foi atribuída uma ruptura numa fronteira em que mais de 2/3 dos avaliadores o fizeram.

Nas 1715 localizações possíveis dos 90 enunciados, o fraseamento de referência atribuiu 448 rupturas, o sistema automático 389 e os avaliadores 370, o que corresponde a uma média de 3.8, 4.4, e 4.6 palavras por constituinte entoacional. Das 389 rupturas atribuídas pelo sistema automático, 26 (6.7%) foram consideradas como falsas inserções, e 30 (7.7%) como apagamentos, o que corresponde a uma taxa de erro de 1.5% e 1.7%, respectivamente, se o desempenho for quantificado, como é habitual, relativamente ao número total de fronteiras do *corpus*. No caso do fraseamento de referência, 15 rupturas (3.3%) foram consideradas falsas inserções e 5 (1.1%) apagamentos, o que corresponde a uma taxa de erro de 0.9% e 0.3%, respectivamente, em relação ao número total de fronteiras. Como a distribuição desses erros mostra (cf. quadro 4), as segmentações de referência e as preditas pelo sistema automático apenas coincidem com as asseguradas pelos avaliadores no máximo em 85% e 69% dos casos, respectivamente. Tendo estes considerado correctas 93% das primeiras e 77% das segundas, torna-se claro que nem todas as diferenças observadas têm igual importância e correspondem a verdadeiros erros.

número de erros	Falsas inserções		Apagamentos	
	automático	referência	Automático	referência
0	74%	84%	69%	94%
1	22%	14%	29%	6%
2	3%	1%	2%	0%

Quadro 4: Número de erros por enunciado e percentagem de ocorrência dos diferentes tipos de erros na totalidade do *corpus* de teste.

A análise cruzada dos resultados das diferentes tarefas, confirma que alguns enunciados onde foram detectados um ou mais erros na atribuição de rupturas podem ser considerados aceitáveis ou mesmo bons. Mas, se apenas forem tidos em consideração os 20 julgados inaceitáveis pela maioria dos avaliadores, verifica-se, que 8 (40%) foram-no devido a erros de apagamento, 6 (30%) devido a falsas inserções e os restantes 6 (30%) por ambas as razões. Estas observações reforçam a hipótese de que tanto a ausência como a presença de uma só ruptura numa fronteira crítica são suficientes para que um enunciado seja rejeitado como um todo.

Essa análise permite ainda identificar os contextos em que a presença/ausência de uma ruptura é crucial, estando alguns dos casos mais problemáticos relacionados com a inserção destas antes de frases relativas restritivas, de constituintes coordenados e de constituintes preposicionais. No primeiro caso, essas rupturas tanto podem conduzir a uma maioria de julgamentos desfavoráveis, como ser consideradas aceitáveis ou mesmo boas se a sua extensão for suficientemente longa e/ou se for necessário desfazer assimetrias rítmicas. Estes factores, que não são independentes um do outro, permitem também explicar parte da variabilidade observada nos outros dois casos.

5. Conclusões

Para lidar com um conjunto tão heterogéneo de dados, como é o deste teste, foi crucial a definição de critérios rigorosos de avaliação que, exigindo uma maioria de opiniões expressas para validar um julgamento, permitem pôr em evidência regularidades no comportamento dos avaliadores e apontam para a necessidade de novas medidas de desempenho que não tenham em consideração apenas o número de rupturas correctamente atribuídas.

A metodologia utilizada permitiu verificar que, embora seja admitida uma grande variabilidade na localização de marcas de ruptura, um enunciado pode ser considerado inaceitável devido a uma única falha, que tanto pode corresponder a uma falsa inserção como a um apagamento. Os dois tipos de erro parecem ser igualmente importantes, o que põe em causa critérios de optimização, frequentemente utilizados, que atribuem diferentes pesos a cada um deles.

O cruzamento dos resultados relativos aos dois tipos de tarefas exigidas neste teste (avaliação e proposta de segmentação) forneceu importantes indicações sobre contextos em que a presença/ausência de uma ruptura é crítica e permitiu avaliar a qualidade do fraseamento prosódico de referência utilizado para treino do modelo e a do fraseamento prosódico automático por ele realizado. Apesar de o fraseamento de referência não apresentar uma taxa de aceitabilidade de 100%, o *corpus* de treino pode considerar-se amplamente validado uma vez que 98.8% das fronteiras de palavra e 93% dos enunciados foram classificados como correctos. O sistema automático, cujo desempenho se pretende ainda melhorar, trata correctamente 96.8% das fronteiras de palavra. Ao nível do enunciado, em que apenas concordava com o fraseamento de referência em 22% dos casos, obteve uma taxa de aceitabilidade de 78%, o que torna evidente a capacidade de generalização do modelo.

Referências

Hirschberg, J.e P. Prieto (1996). "Training intonation phrase rules automatically for English and Spanish texto-to-speech". *Speech Communication*, 18: 281-290.

Huang, X., A. Acero e H. Hon (2001). *A guide to Theory, Algorithm, and System Development*. Prentice Hall.

Taylor, P. e A. W. Black (1998). "Assigning Phrase Breaks from Part-of-Speech Sequencies". *Computer, Speech and Language*, 12 (2): 99-117.

Viana, M.C. e A. I. Mata (2001). "Fraseamento prosódico em PE com técnicas CART".In *Razões e Emoção* Miscelânea de estudos oferecida a Maria Helena Mateus, Departamento Linguística Geral e Românica, FLUL Lisboa, Junho de 2001.

Agradecimentos

Este estudo foi desenvolvido no âmbito do projecto projecto *Dixi+: Sistema de Síntese para Comunicação Alternativa e Aumentativa*", financiado pela Fundação para a Ciência e Tecnologia (FCT), Programa PRAXIS XXI. Gostariamos de agradecer, em particular, à Prof. Isabel Trancoso pelas suas críticas e sugestões que em muito contribuíram para o desenvolvimento deste trabalho.