Discriminação automática de sinais de pontuação e de disfluências

Helena Moniz[1,2], Fernando Batista[2,3], Ana I. Mata[1] e Isabel Trancoso[2,4]

[1] Faculdade de Letras da Universidade de Lisboa
[2] INESC-ID, Lisboa
[3] ISCTE-IUL - Instituto Universitário de Lisboa
[4] Instituto Superior Técnico, Universidade de Lisboa
{helena.moniz; fernando.batista; isabel.trancoso}@inesc-id.pt e aim@fl.ul.pt

Abstract:
This paper discriminates different types of structural metadata, namely punctuation marks and disfluencies, in transcripts of university lectures. The disambiguation process is based on predefined multilayered linguistic information for those boundary events. Since boundary events may share similar linguistic properties, in terms of f0 and energy slopes, presence/absence of silent pauses, and duration of different units of analysis, distinct classification methods based on a set of automatically derived prosodic features have been applied to differentiate between punctuation marks and disfluencies. This paper also performs a detailed analysis on the impact of each individual features in discriminating each structural metadata event.

Keywords: structural metadata events, disfluencies, punctuation, speech prosody, automatic speech processing.

Palavras-chave: disfluências, marcas de pontuação, prosódia, processamento automático de fala.

1 Introdução

O presente trabalho tem como objectivo classificar automaticamente marcas de pontuação e disfluências, com base em pistas linguísticas, de natureza essencialmente prosódica. Pretende-se contribuir para o processamento automático de eventos designados no inglês "structural metadata events", *i. e.*, a recuperação automática de pontuação e maiúsculas em fronteiras de frase, bem como a anotação e filtragem de disfluências (*e.g.*, Liu *et al*., 2006; Ostendorf *et al.*, 2008).

O enriquecimento automático de transcrições de fala com identificação de marcas de pontuação e disfluências contribui significativamente para a legibilidade de uma sequência de palavras obtida através de um reconhecedor automático de fala. Este processo de enriquecimento de transcrições é de tal forma crucial para diversas aplicações que frequentemente o reconhecedor de fala está integrado numa sequência de módulos que incluem, por exemplo, segmentar áudio, grafar a maiúsculas e minúsculas, identificar zonas de disfluência. A tarefa de enriquecimento de transcrições de fala pode ser entendida como uma tarefa de estruturação de uma sequência de palavras em diferentes unidades linguísticas, ou seja, uma estruturação multi-nível que integra distintos módulos da gramática.

Para o processo de enriquecimento de transcrições de fala têm sido utilizadas distintas pistas, reflexo da estruturação multi-nível acima referida. Estas pistas estão muito para além das lexicais (n-gramas de palavras) diretamente extraídas da saída do reconhecedor, ou das acústicas extraídas do módulo de processamento de áudio (identificação de segmentos correspondentes a fala vs. outros tipos de segmentos nos quais se incluem silêncios, música, *inter alia*; classificação de falantes, etc.). Saliente-se que de um

***Textos Selecionados, XXIX Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2014, pp. 395-405, ISBN 978-989-97440-3-5**

conjunto alargado de pistas, as prosódicas têm merecido destaque no processo de enriquecimento de transcrições de fala. Por conseguinte, o objectivo deste trabalho centra-se na análise de pistas prosódicas e na aferição do contributo das mesmas para o enriquecimento de transcrições de fala.

O processo de enriquecimento de transcrições é assaz complexo, uma vez que implica pontuar fala, por um lado, e identificar disfluências, por outro – tarefas descritas como factores que afectam sobremaneira a concordância inter-anotadores em distintos corpora do português europeu (Batista, 2011 e Cabarrão *et al.*, 2014 relativamente a notícias televisivas; Moniz, 2013 relativamente a aulas universitárias). Acresce que as pistas prosódicas quer para a atribuição de uma marca de pontuação quer para a identificação de uma sequência disfluente podem ser ambíguas, como já descrito em Batista *et al.* 2012 e Moniz *et al.* 2012, visto tratarem-se de eventos maioritariamente produzidos em fronteiras de constituintes prosódicos.

2 Estado de arte

O impacto de distintos métodos de aprendizagem automática e de pistas prosódicas tem sido discutido na literatura crítica sobre enriquecimento de transcrições de fala. Shriberg et al (2000) e Christensen *et al.* (2001) aplicaram Modelos de Markov Observáveis e combinaram pistas lexicais e prosódicas para a recuperação de pontos finais, vírgulas e pontos de interrogação. Por sua vez, Huang & Zweig (2002) utilizaram um modelo de Máxima Entropia também para a recuperação de pontos finais, vírgulas e pontos de interrogação, sendo que os melhores resultados são conseguidos com a combinação de pistas lexicais e prosódicas. Wang & Narayanan (2004) propuseram um algoritmo para a segmentação de frases tendo por base a relação prosódica entre índices de ruptura e duração de várias unidades linguísticas. Liu *et al.* (2006), Ostendorf *et al.* (2008) e Favre *et al.* (2009) mostraram, com base em diferentes métodos, que a recuperação de marcas de pontuação deve abarcar a combinação de pistas prosódicas, morfológicas e sintácticas.

Tanto os métodos como as pistas linguísticas utilizados na recuperação de marcas de pontuação são também aplicados na identificação de disfluências. O que é distinto relativamente ao último caso é o facto de as disfluências terem uma estruturação idiossincrática que compreende: o *reparandum*, o momento de interrupção, o *interregnum* e a reposição da fluência (Levelt, 1989; Nakatani & Hirschberg, 1994; Shriberg, 1994). O *reparandum* corresponde à zona a ser corrigida. O ponto de interrupção corresponde ao momento em que o falante interrompe o seu discurso para corrigir o material linguístico, pode ser considerado como a fronteira entre o discurso disfluente e o fluente. O *interregnum* é um intervalo que pode conter pausas preenchidas (*aa*, *aam*, *mm*, *ee*, *eem*), silenciosas ou marcadores de edição (*e.g.*, *quer dizer*, *ou melhor*, *não é isto*). Finalmente, a última zona corresponde à reposição da fluência. É sabido que cada uma das regiões de uma sequência disfluente tem propriedades acústicas específicas e diferenciadoras das distintas zonas, como preconizado na teoria da Edição do Sinal definida por Hindle (1983), *i. e.*, os falantes sinalizam que estão a corrigir e a repor a fluência. A edição do sinal é efectuado com base em diversas pistas. O *reparandum* é sobretudo editado através da produção de fragmentos, glotalizações, gestos co-articulatórios e atributos de qualidade de voz característicos, tal como perturbações nos períodos de f0 (jitter). O interregnum é editado com base em durações de pausas significativamente distintas e na selecção lexical de marcadores de edição e/ou pausas preenchidas. A reposição da fluência caracteriza-se por padrões prosódicos de marcação por contraste ou de paralelismo.

Relativamente aos métodos utilizados na detecção de disfluências, Heeman & Allen (1999) descreveram um modelo de língua estatístico que integrava a identificação de categorias morfológicas, marcadores discursivos, disfluências e constituintes prosódicos tratados simultaneamente, demonstrando que os melhores resultados foram obtidos com a análise simultânea de todos os eventos. Nakatani & Hirschberg (1994) bem como Shriberg (1999) utilizaram Árvores de Decisão e Regressão para a identificação dos pontos de interrupção com base em pistas prosódicas. Kim & Woodland (2004) e Liu et al. (2006) combinaram pistas prosódicas e lexicais para classificarem disfluências e marcas de pontuação.

Este trabalho classifica tanto marcas de pontuação como disfluências com base num conjunto de pistas prosódicas reportadas nos estudos mencionados, nomeadamente, presença e duração de pausas nas fronteiras dos eventos, declinação global de f0, reinicializações de f0 e de energia, alongamento final dos segmentos, ocorrência de pausas preenchidas e ocorrência de fragmentos. Espera-se que esta análise para

o português europeu possa contribuir para questões de investigação ainda em aberto relativas ao impacto de pistas linguísticas distintas por tarefas, domínios e línguas.

3 Corpus

A classificação das marcas de pontuação e de disfluências foi realizada com base no corpus LECTRA (Trancoso *et al.*, 2008), um corpus de aulas universitárias com aproximadamente 31 horas anotadas a vários níveis. O corpus foi dividido em subconjuntos de treino+desenvolvimento (89%) e teste (11%), divididos em função de critérios temporais, ou seja, as primeiras aulas de cada curso foram incluídas no subconjunto de treino e as últimas nos de desenvolvimento e teste.

	Treino/desenvolvimento	**Teste**
Tempo (horas)	28:00	3:24
Número de palavras	216435	24516
Número de pausas preenchidas	8390	950
Número de reposições de fluência	5608	720
Número de pontos finais	8363	861
Número de vírgulas	22957	2612
Número de pontos de interrogação	3526	498

Quadro 1 - Propriedades do corpus.

Sobre a distribuição das marcas de pontuação e de disfluência/reposição de fluência no corpus em questão, *vide* Quadro 1. Da leitura do quadro verifica-se que a frequência da vírgula é bastante superior a qualquer outro evento, perfazendo mesmo 50% de todos os eventos no corpus. Num estudo comparativo entre o português europeu e o inglês americano em corpora de notícias televisivas, Batista *et al.* (2012) verificaram que a percentagem de vírgulas no português corresponde ao dobro das do inglês.

4 Reconhecimento automático e integração de pistas prosódicas

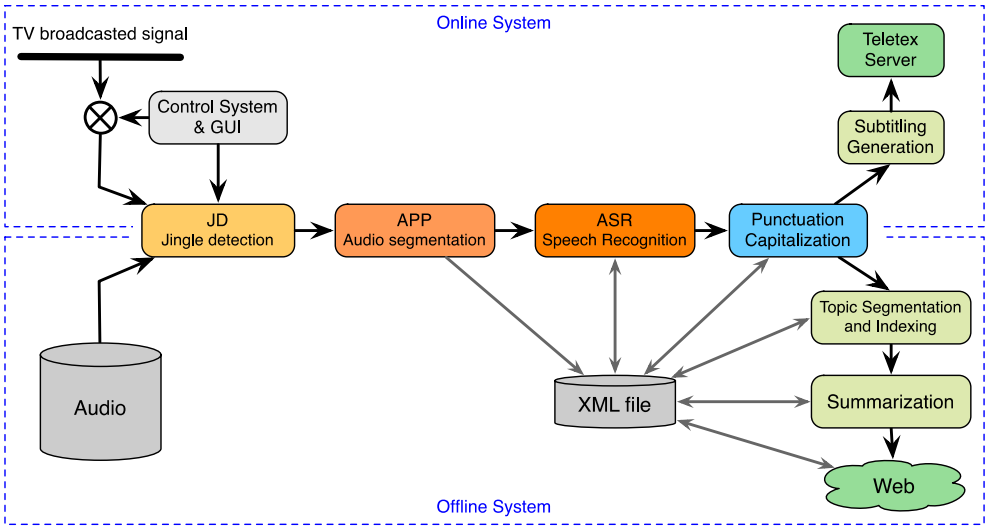

Figura 1 –Arquitetura do reconhecedor automático de fala.

O reconhecedor automático de fala corresponde a uma arquitetura em módulos, especificamente construída para notícias televisivas, como ilustrado na Figura 1. Num primeiro momento, os áudios dos

noticiários televisivos são gravados (*Control System*). Posteriormente, o módulo de detecção de *jingles (Jingle Detection)* detecta o início e o final do noticiário, bem como potenciais intervalos. Seguidamente, os segmentos anteriormente classificados como fala são pré-processados (*APP, Audio Pre-Processing or Audio Segmentation*), seguindo o modelo de Meinedo & Neto (2008), ou seja, são identificados os falantes e o respectivo género e são classificadas as condições acústicas (fala limpa, ruído, etc.). O reconhecedor Audimus (Neto *et al*., 2003; Neto *et al.* 2008; Meinedo & Neto, 2008), um módulo de reconhecimento de fala contínua, processa cada segmento de fala previamente identificado no módulo de segmentação de áudio e produz uma transcrição inicial. Os módulos de pontuação e de maiusculização (Batista, 2011) processam as transcrições originais, enriquecendo-as com recuperação de marcas de pontuação. O sistema tem vindo a ser utilizado para legendagem automática no modo *on-line* ou/e para produção de conteúdos multimédia num modo *offline*. A versão *on-line* utiliza o gerador de legendas (Subtitling Generator), culminando no envio da legenda para o servidor de teletextos (Teletex Server). A versão *offline* compreende o enriquecimento de um ficheiro XML, contendo a informação organizada dos módulos de segmentação de áudio e de reconhecimento, com segmentação e indexação de tópicos (*Topic Segmentation and Indexing*), proposta por Amaral & Trancoso (2008), e de sumários das notícias (Ribeiro & Matos, 2011). Os conteúdos são posteriormente disponibilizados na Web.

No trabalho aqui reportado, o reconhecedor acima descrito foi usado no modo de alinhamento forçado e esta opção deveu-se a duas razões principais. Primeiro, trata-se de um reconhecedor treinado com dados de notícias televisivas, um domínio bastante distinto do das aulas universitárias. Segundo, a escassez de textos escritos em português sobre os conteúdos das matérias leccionadas não permite treinar modelos de língua apropriados para este domínio. O recurso ao alinhamento forçado permite não influenciar negativamente os resultados obtidos com um reconhecedor não adaptado ao domínio particular em análise. Embora o alinhamento forçado permita uma taxa de acerto bastante superior à de alinhamento automático, ainda persiste 0.9% de taxa de erro no alinhamento de palavras, sobretudo devido a regiões com baixa energia.

Para o tratamento automático de marcas de pontuação e de disfluências concorrem diversas fontes de informação, designadamente, as transcrições manuais, as produzidas pelo reconhecedor e o sinal acústico (para mais informações sobre este processo, veja-se Batista *et al.*, 2012b). Após realizado o reconhecimento automático, as anotações manuais, nas quais se incluem as marcas de pontuação e as zonas de disfluências, são adicionadas às transcrições automáticas, através da ferramenta NIST SCLite (http://www.ist.gov/speech). A Figura 1 apresenta um excerto ilustrativo da informação fornecida pelo reconhecedor acrescida de informação sobre a identificação de palavras (dis)fluentes: a indicação dos tempos de início e fim de cada palavra; a confiança da palavra; as condições de fala ("F1" para fala sem ruído); a categoria morfológica e a identificação da palavra propriamente dita.

```
<TranscriptSegment start="3694" end="4316">
  <TranscriptWordList type="disfluency" start="3694" end="3753">
    </Word start="3694" end="3753" conf="0.996" focus="F1" pos="S." name="na">
  </TranscriptWordList>
  <TranscriptWordList type="chunk" start="3754" end="3919">
    </Word start="3754" end="3791" conf="0.999" focus="F1" pos="A." name="última">
    </Word start="3792" end="3824" conf="0.998" focus="F1" pos="Nc" name="aula">
    </Word start="3829" end="3865" conf="0.997" focus="F1" pos="V." name="estávamos">
    </Word start="3866" end="3869" conf="0.997" focus="F1" pos="S." name="a">
    </Word start="3870" end="3919" conf="0.999" focus="F1" pos="V." name="falar">
  </TranscriptWordList>
  <TranscriptWordList type="disfluency" start="3943" end="3975">
    </Word start="3944" end="3975" conf="0.685" status="filled_pause" focus="F1"
   name="%aa">
  </TranscriptWordList>
  <TranscriptWordList type="chunk" start="3976" end="4316">
    </Word start="3976" end="3989" conf="0.997" focus="F1" pos="S." name="de">
    </Word start="3990" end="4039" conf="0.282" focus="F1" pos="Nc" name="bases">
    </Word start="4213" end="4230" conf="0.998" focus="F1" pos="S." name="de">
    </Word start="4231" end="4269" conf="0.999" focus="F1" pos="Nc" name="espaços">
    </Word start="4270" end="4316" conf="0.999" focus="F1" punct="." pos="A."
   name="lineares">
  </TranscriptWordList>
</TranscriptSegment>
```

Figura 2 – Excerto com saída do reconhecedor e classificação de palavras (dis)fluentes.

A informação relativa a f0 e energia não estava disponível nos módulos do reconhecedor, pelo que foi necessário extrair a informação referida com a ferramenta de acesso público Snack Sound Toolkit (Sjölander *et al*., 1998). Os declives (do inglês *"slope"*) de f0 e de energia foram calculados com base em Regressão Linear. Pistas acústicas segmentais e suprassegmentais foram extraídas automaticamente e organizadas numa estruturação hierárquica que contempla fones, sílabas, palavras, unidades similares a frases (do inglês *sentence like-units*) e actos de fala. A Figura 3 ilustra um excerto com as pistas relativas a f0 e energia. Valores mínimos, máximos, médias, medianas, desvios padrão e declives foram extraídos para cada fone, sílaba e palavra.

```
<Word start="5053" end="5099" conf="0.999813" focus="F3" punct="." pos="Nc" name="noite"
      phseq="_noj#t@+" pmax="268.3" pmin="74.8" pavg="212.3" pmed="209.1" pstdev="34.57" emax="62.8"
      emin="32.9" eavg="52.1" emed="58.2" estdev="10.17" eslope="-0.3" eslope_norm="-12.14" pslope="-0.18"
      pslope_norm="-8.41" pmin_st_100="-5.03" pmax_st_100="17.09" pmin_st_spkr="1.70"
      pmax_st_spkr="23.81" pmin-zscore_spkr="-3.25">
 <syl stress="y" start="5053" dur="25.5" pmax="268.3" pmin="196.2" pavg="218.0" pmed="210.2" pstdev="20.60"
      emax="62.8" emin="37.9" eavg="56.9" emed="59.3" estdev="5.98" eslope="-0.2" eslope_norm="-4.89"
      pslope="0.17" pslope_norm="4.08">
  <ph name="n" start="5053" dur="7" pmax="216.2" pmin="201.0" pavg="208.5" pmed="209.1" pstdev="4.41"
      emax="60.1" emin="52.9" eavg="55.7" emed="54.6" estdev="2.78"/>
  <ph name="o" start="5060" dur="9" pmax="215.3" pmin="196.2" pavg="203.0" pmed="200.1" pstdev="6.37"
      emax="60.5" emin="58.5" eavg="59.5" emed="59.5" estdev="0.63"/>
  <ph name="j" start="5069" dur="9.5" pmax="268.3" pmin="221.2" pavg="243.3" pmed="241.1" pstdev="15.49"
      emax="62.8" emin="37.9" eavg="55.4" emed="60.3" estdev="8.81"/>
 </syl>
 <syl start="5078.5" dur="21.5" pmax="74.8" pmin="74.8" pavg="74.8" pmed="74.8" pstdev="0.00" emax="60.9"
      emin="32.9" eavg="45.9" emed="40.6" estdev="11.03" eslope="1.1" eslope_norm="23.29" pslope="0.00"
      pslope_norm="0.00">
  <ph name="t" start="5078.5" dur="8.5" emax="40.0" emin="32.9" eavg="35.8" emed="35.1" estdev="2.22"/>
  <ph name="@" start="5087" dur="13" pmax="74.8" pmin="74.8" pavg="74.8" pmed="74.8" pstdev="0.00"
      emax="60.9" emin="33.2" eavg="52.8" emed="58.0" estdev="9.12"/>
 </syl>
</Word>
```

Figura 3 – Excerto do ficheiro XML contendo a informação prosódica para a palavra "noite".

5 Predição de marcas de pontuação e disfluências

As experiências realizadas no âmbito do presente trabalho têm por base um conjunto de pistas acústicas puramente automáticas, extraídas da saída do reconhecedor e do sinal acústico, como previamente descrito na secção anterior. As pistas são extraídas dos contextos adjacentes ao evento, particularmente das duas palavras imediatamente anteriores e da palavra seguinte ao evento. As pistas que envolvem apenas uma palavra compreendem: declives de f0 e de energia, níveis de confiança atribuídos pelo reconhecedor, duração da palavra, número de fones e de sílabas. As pistas que envolvem duas palavras consecutivas incluem: contornos de f0 e de energia, diferenças de f0 e de energia entre as duas palavras, presença/ausência de silêncios, duração relativa de silêncios adjacentes, durações das palavras, medianas de f0 e de energia.

As experiências foram realizadas com recurso à aplicação de acesso público Weka (Hall *et al.*, 2009). Diferentes métodos de aprendizagem estatística foram aplicados: *Naïve Bayes*, Regressão Linear e Árvores de Decisão e Regressão (CART). Os resultados foram avaliados de acordo com as métricas-padrão (Makhoul *et al.*, 1999), designadamente, as de precisão (o número de eventos correctos "tp" sobre o número de eventos na hipótese), cobertura (o número de eventos correctos sobre o número total de eventos na referência), medida-F e *Slot Error Rate* (SER, corresponde a uma taxa de erro, mais concretamente ao número de erros na identificação dos eventos dividido pelo número de eventos existentes na referência).

$$precisão = \frac{tp}{tp+fp}$$

$$cobertura = \frac{tp}{tp+fn}$$

$$medida - F = \frac{2 \times precisão \times cobertura}{precisão + cobertura}$$

Os melhores resultados da predição de marcas de pontuação e de disfluências foram consistentemente obtidos com as CART. Os resultados serão discutidos na secção 4.1. e a análise das pistas mais informativas será apresentada na 4.2.

5.1 Resultados

Para a discriminação entre marcas de pontuação e disfluências foram consideradas quatro classes: ponto final, vírgula, ponto de interrogação e disfluências. O Quadro 2 apresenta os resultados obtidos com as CART. A classe mais bem classificada corresponde ao ponto final, tal como era expectável, uma vez que é sabido para o português (Batista *et al.*, 2012) que as pistas prosódicas são cruciais para a detecção desta classe. Relativamente ao evento menos bem predito, as vírgulas são parcamente identificadas com base em pistas prosódicas, resultados em concordância com estudos realizados para outras línguas (artigo considerado estado e arte, Favre *et al.*, 2009) nos quais as vírgulas são sobretudo classificadas com base em pistas morfo-sintácticas. Os pontos de interrogação são apontadas na literatura como sendo uma classe difícil de detectar (*e.g.*, Moniz *et al.*, 2011; Margolis, 2011) por corresponderem a distintos padrões prosódicos em função dos tipos de interrogativas. Acresce-se que as interrogativas em português compreendem um menor número de pistas lexicais do que, por exemplo, em inglês uma vez que a inversão sujeito verbo ("do you") permite uma predição de interrogativas com taxas superiores de sucesso no inglês. A classe com resultados inferiores corresponde às reposições de fluência. Embora seja conhecido para o português que as reposições de fluência são caracterizadas por padrões prosódicos de marcação de contraste (Moniz *et al.*, 2012; Moniz, 2013), numa tarefa de predição de marcas de pontuação e de disfluências, estas não são na sua maioria identificadas, *vide* matriz de confusão no Quadro 3. Uma possível explicação para este facto poderá ser a de as palavras correspondentes a

reposições de fluência serem confundidas com palavras produzidas após uma marca de pontuação que não correspondem a reposições de fluência, mas partilham com estas a reinicialização de f0 e de energia.

Classes	**Precisão**	**Cobertura**	**Medida-F**	**SER**
vírgula (,)	60.6	27.6	37.9	90.3
ponto final (.)	64.1	67.6	65.8	70.2
ponto de interrogação (?)	73.9	29.5	42.2	80.9
reposição de fluência	60.8	13.1	21.6	95.4
média ponderada	63.0	32.9	43.3	75.6

Quadro 2- Resultados obtidos com CART, com base em pistas prosódicas.

Classificados como:					
	,	.	?	reposição	apagamentos
vírgula (,)	718	36	10	15	1823
ponto final (.)	76	579	35	3	163
ponto de interrogação (?)	27	225	147	4	95
reposição de fluência	51	19	1	93	546
inserções	312	44	6	38	

Quadro 3 - Matriz de confusão resultante da classificação dos eventos.

Tem sido apontado na literatura (*e.g.*, Levelt, 1989; Nakatani & Hirschberg, 1994; Shriberg, 1994) que fragmentos e pausas preenchidas podem ser utilizados como pistas para detectar as zonas estruturantes de uma disfluência. Suportados nos estudos referidos, realizou-se uma experiência em que se acrescentou pausas preenchidas e fragmentos como pistas. Os resultados mostram que a utilização destas duas categorias como pistas melhora o desempenho da classificação das reposições de fluência para 48.8% medida-F e a para um valor de classificação global de 47.8%. Constata-se ainda que o impacto relativamente ao uso de fragmentos como pista é bastante inferior ao reportado por Nakatani & Hirschberg (1994) ou Shriberg (1994), uma vez que os fragmentos nas aulas universitárias correspondem somente a 6.6% de todos os tipos de disfluências.

5.2 Análise das pistas mais informativas

	Pista	nada	vírgula	ponto final	?	reposição de fluência
1	declives de f0: F- (pw,cw)			***	*****	
2	declives de f0: -- (pw,cw)			****	*****	
3	declives de f0: R- (pw,cw)			****	*****	
4	confiança da palavra (cw)			*****	****	
5	declives de energia: RF (cw,fw)			*****	****	
6	declives de energia: -- (pw,cw)			*****	**	
7	declives de energia: F- (pw,cw)			*****	***	
8	declives de energia: R- (cw,fw)			*****	***	
9	declives de energia: R- (pw,cw)			*****	**	
10	declives de energia: RF (pw,cw)			***	*****	
11	declives de energia: FF (pw,cw)			***	*****	
12	declives de energia: RR (cw,fw)			****	****	
13	declives de energia: -F (pw,cw)			****	****	
14	declives de f0: RF (cw,fw)		*	*	*****	
15	declives de f0: F- (cw,fw)			*****	***	
16	declives de f0: FF (cw,fw)			*****	***	
17	declives de f0: R- (cw,fw)		*	*	*****	
18	declives de f0: RR (cw,fw)		*	*	*****	
19	declives de f0: FR (cw,fw)			*****	***	
20	relação entre silêncio anterior (cw,fw)	*****				
21	comparação entre os silêncios anteriores: > (cw,fw)			***	****	
22	relação entre a mediana da energia (cw,fw)	*	*	***	***	
23	relação entre silêncio anterior (pw,cw)	*****	**			*
24	duração.ratio (cw,fw)		*	*****		*
25	relação entre a duração (pw,cw)	***	**	*		*
26	relação entre a mediana da energia (pw,cw)	**	*	***		*
27	declives de f0: -F (pw,cw)	****	*			***
28	declives de f0: RF (pw,cw)	****	**			***
29	declives de f0: FF (pw,cw)	****	**			***
30	declives de f0: -F (cw,fw)	****				*****
31	declives de energia: -F (cw,fw)	***	**			***
32	declives de f0: -- (cw,fw)	***				*****
33	palavras iguais (pw,cw)	*	*	***		**
34	declives de f0: -R (cw,fw)	***				*****
35	fones (cw)	*	**	**	**	*
36	comparação entre os silêncios anteriores: < (cw,fw)	***	**			**
37	comparação entre os silêncios anteriores: > (pw,cw)	*	*	**	***	*
38	declives de energia: -R (cw,fw)	**	**			****
39	declives de energia: -- (cw,fw)	**	**			****
40	relação entre a mediana do f0 (pw,cw)	**	*	**	*	*
41	declives de energia: FR (cw,fw)		**			*****
42	declives de f0: -R (pw,cw)	**				*****
43	declives de energia: RR (pw,cw)	*	*			*****
44	declives de energia: -R (pw,cw)	*	**			*****
45	declives de energia: FR (pw,cw)	*	**			*****
46	declives de energia: F- (cw,fw)	*	**			*****
47	declives de f0: FR (pw,cw)	**				*****
48	declives de f0: RR (pw,cw)	**				*****
49	palavras iguais (cw,fw)		*		*	*****
50	declives de energia: FF (cw,fw)	*	**			*****
51	confiança (fw)	*	*		*	*****
52	comparação entre os silêncios anteriores: = (cw,fw)	**	*	*	*	**
53	comparação entre os silêncios anteriores: = (pw,cw)	**	**	*	*	**

Quadro 4 – Pistas mais relevantes na discriminação entre marcas de pontuação e disfluências

O Quadro 4 mostra as pistas mais informativas na discriminação entre marcas de pontuação e disfluências ordenadas por relevância, em que *pw* corresponde à palavra anterior, *cw* à palavra atual e *fw* à palavra seguinte. Os asteriscos indicam a relevância da pista, sendo que um maior número de asteriscos corresponde a uma maior relevância. Para a discriminação destas classes de eventos é determinante o seguinte conjunto de pistas linguísticas: contornos de frequência fundamental (f0), níveis de energia, duração relativa das unidades de análise e grau de confiança dessas mesmas unidades.

Em primeiro lugar, as pistas que mais contribuem para a predição da reposição de fluência a seguir a uma sequência disfluente integram: i) duas palavras contíguas idênticas; ii) subida dos níveis de f0 e de energia na palavra que inicia uma reposição de fluência e um contorno estacionário de f0 na palavra anterior; iii) grau de confiança da palavra que inicia a reposição, superior ao da disfluência propriamente dita.

Relativamente às pistas associadas à predição de pontos finais, estas incluem: i) contorno descendente na palavra antes de um ponto final; ii) nível estacionário de energia na mesma palavra; iii) duração relativa entre essa palavra e a seguinte; e iv) grau superior de confiança em relação à palavra seguinte. Este conjunto de pistas é ilustrativo do comportamento de uma declarativa neutra no PE.

Os pontos de interrogação, por sua vez, são caracterizados por dois padrões diferenciados: i) contorno de f0 ascendente na palavra antes de um ponto de interrogação e declive (do inglês "slope") de energia ascendente nessa e na palavra seguinte; ii) contorno de f0 estacionário na palavra antes de um ponto de interrogação e declive de energia descendente nessa mesma palavra.

As vírgulas são o evento que menos depende de uma caracterização prosódica. Nas experiências até agora realizadas para o PE, elas são sobretudo classificadas com base em pistas morfo-sintácticas, não sendo claramente desambiguadas por meio de pistas prosódicas.

No que diz respeito às palavras em fala contínua que não contêm marcas de pontuação e que não são disfluentes, a pista mais informativa é a ausência de silêncios, como expectável.

6 Conclusões

Este trabalho descreve experiências de classificação de marcas de pontuação e de disfluências num corpus de aulas universitárias, um domínio caracterizado por uma elevada percentagem dos referidos eventos. Estudos prévios realizados para o português europeu mostram que i) a recuperação de marcas de pontuação está estreitamente correlacionada com a selecção de conjuntos de pistas linguísticas e ii) diferentes tipos de frases influenciam a relevância de determinadas pistas em detrimento de outras. Partindo destes estudos optou-se por fazer uma análise que discriminasse as diferentes marcas de pontuação e que salientasse as pistas mais informativas para cada uma destas marcas. Procurou-se também desambiguar as marcas de pontuação das disfluências, uma vez que, como eventos produzidos em fronteiras de constituintes prosódicos, partilham propriedades linguísticas. Com base num conjunto de pistas puramente prosódicas, foi possível predizer padrões regulares na identificação das distintas marcas de pontuação e de reposição de fluência. O conjunto de experiências efectuado no âmbito deste trabalho constitui-se como um primeiro contributo para a sistematização das propriedades linguísticas associadas a sinais de pontuação e a reposição de fluência em PE.

O trabalho futuro procurará alargar o estudo da recuperação de marcas de pontuação e de reposição de fluência ao reconhecimento automático (no estudo apresentado utilizou-se o reconhecedor em modo de alinhamento forçado) com base na integração de pistas morfo-sintácticas.

Agradecimentos

Este trabalho foi financiado com fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, pela bolsa de Pós-doutoramento SFRH/BPD/95849/2013 e pelos projetos PTDC/CLE-LIN/120017/2010 (COPAS) e PEst-OE/EEI/LA0021/2013.

Referências

Batista, F., Moniz, H., Trancoso, I. & Mamede, N. (2012a) Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts. IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, n. 2, pp. 474 – 485.

Batista, F., Moniz, H., Trancoso, I., Mamede, N., & Mata, A.I. (2012b) Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. Journal of Speech Sciences, (3).

Batista, F. (2011) Recovering capitalization and punctuation marks in speech transcripts. Dissertação de Doutoramento, Instituto Superior Técnico.

Boakye, K., Favre, B., & Hakkani-Tür, D. (2009) Any questions? Automatic question detection in meetings. ASRU 2009, Merano, Itália.

Cabarrão, V., Moniz, H., Batista, F., Ribeiro, R., Mamede, N., Meinedo, H., Trancoso, I., Mata, A. I. & Matos, D. (2014) Revising a broadcast news corpus – a linguistic approach. LREC 2014, Islândia.

Christensen, H., Gotoh, Y. & Renals, S. (2001) Punctuation anntation using statistical prosody models. ASRU 2001, Itália.

Duarte, I. (2000) Instrumentos de análise. Universidade Aberta.

Favre, B., Hakkani-Tür, D. & Shriberg, E. (2009) Syntactically informed models for comma prediction. ICASSP 2009, Taipei, Taiwan.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009) The weka data mining software: an update. SIGKDD Explorer Newsletter, vol. 11, n. 1, pp. 10-18.

Heeman, P. and Allen, J. (1999) Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue. Computational Linguistics, vol. 25, pp. 527–571.

Hindle, D. (1983) Deterministic parsing of syntactic non-fluencies. ACL-83, pp. 123-128.

Huang, J. & Zweig, G. (2002) Maximum entropy model for punctuation annotation of speech. ICSLP 2002, Denver, U.S.A.

Kim, J., Schwarm, S., & Ostendorf, M. (2004) Detecting structural metadata with decision trees and transformation-based learning. HLT-NAACL 2004, pp. 137–144.

Levelt, W. (1989) Speaking. MIT Press, Cambridge, Massachusetts.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M. & Harper, M. (2006) Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on Audio, Speech and Language Processing, 14(5): 1526-1540.

Makhoul, J., Kubala, F., Schwartz R., & Weischedel, R. (1999) Performance measures for information extraction. DARPA Broadcast News Workshop, pp. 249-252.

Margolis, A. (2011). Automatic annotation of spoken language using out-of-domain resources and domain adaptation. Dissertação de doutoramento, Universidade de Washington.

Meinedo, H., Viveiros, M. & Neto, J. (2008) Evaluation of a live broadcast news subtitling system for Portuguese. Interspeech 2008, Brisbane, Austrália.

Moniz, H., 2013. Processing disfluencies in European Portuguese. Dissertação de doutoramento, Universidade de Lisboa.

Moniz, H., Batista, F., Trancoso, I. & Mata, A. I. (2012) Prosodic context-based analysis of disfluencies. Interspeech 2012, Portland, E.U.A.

Moniz, H., Batista, F., Trancoso, I. & Mata, A. I. (2011). Analysis of interrogatives in different domains. In A. Esposito, A. M. Esposito, F. Martone, V. Muller & G. Scarpetta (eds.) Towards autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues. Springer/Heidelberg, Lecture Notes on Computer Science, pp. 136-148, Caserta, Itália.

Nakatani, C. & Hirschberg, J. (1994) A corpus-based study of repair cues in spontaneous speech. Journal of the Acoustical Society of America, vol. 95, pp. 1603–1616.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., & Caseiro, D. (2008) Broadcast news subtitling system in Portuguese. ICASSP'08, pp. 1561–1564.

Neto, J., Meinedo, H., Amaral, R. & Trancoso, I. (2003) The development of an automatic system for selective dissemination of multimedia information. International Workshop on Content-Based Mulimedia Indexing, Rennes, França.

Ostendorf, M. Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J. G., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., & Wooters, C. (2008) Speech segmentation and spoken document processing. IEEE Signal Processing Magazine, 25(3), pp. 59-69.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000) Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication, 32(1-2), pp. 127–154.

Shriberg, E. (1999) Phonetic consequences of speech disfluency. ICPhS'99, São Francisco, U.S.A., pp. 612–622.

Shriberg, E. (1994) Preliminaries to a theory of speech disfluencies.Dissertação de doutoramento, Universidade da California.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998) Web-based educational tools for speech technology. ICSLP 1998, pp. 3217–3220, Sydney, Austrália.

Trancoso, I., Martins, R., Moniz, H., Mata, A. I. & Viana, M. C. (2008) The LECTRA corpus: classroom lecture transcriptions in European Portuguese. LREC 2008, Language Resources and Evaluation Conference, Marraquexe, Marrocos.

Wang, D. & Narayanan, S. (2004) A multi-pass linear algorithm for sentence boundary detection using prosodic cues. International Conference on Acoustics, Speech, and Signal Processing. Montreal, Canadá.