# Análise de interrogativas em diferentes domínios

*Helena Moniz[1,2], Fernando Batista[2,3], Isabel Trancoso[2,4] e*
*Ana Isabel Mata[1]*
[1]FLUL/CLUL; [2]L2F/INESC-ID; [3]ISCTE; [4]IST

**Abstract**

The aim of this work is twofold: to quantify the distinct interrogative types in different domains for European Portuguese, and to discuss the weight of the linguistic features that best describe these structures. The statistical analysis confirms that the percentage of the different types of interrogative is highly dependent on the nature of the *corpus*. Experiments on the automatic detection of interrogatives, using only lexical cues, show results that are strongly correlated with the detection of wh- questions. When prosodic features are added, yes/no questions are then increasingly identified, showing the advantages of combining both lexical and prosodic information.

**Keywords/Palavras-chave:** Interrogatives, punctuation, and prosody. Interrogativas, pontuação e prosódia.

## 1. Introdução

A detecção automática de interrogativas tem especial relevo em diferentes aplicações, como a pontuação de dados de fala reconhecidos automaticamente. O módulo da pontuação previamente desenvolvido para o português europeu apenas dava conta dos pontos finais e das vírgulas. Um dos objectivos do corrente trabalho é, precisamente, o de alargar o escopo de análise e dar conta também das interrogativas, através da avaliação das propriedades linguísticas que caracterizam essas estruturas.

No presente estudo, procura-se quantificar os tipos distintos de interrogativas em diferentes domínios no português europeu (PE) e discutir o peso relativo das diversas pistas linguísticas que melhor descrevem as referidas estruturas, para que o seu processamento automático possa ser realizado em dados de fala preparada e espontânea.

Tal como em outras línguas, existem diferentes tipos de interrogativas no PE: totais ou de sim-não, interrogativas Qu- ou parciais e *tags* ou de confirmação (Mateus *et al.*, 2003). Optou-se ainda por considerar as interrogativas alternativas como um tipo específico, para as discriminar das coordenadas disjuntivas. Uma interrogativa de sim--não, tal como o nome indica, implica, tipicamente, uma resposta afirmativa ou negativa ao conteúdo de toda a proposição (*e.g.*, *Estão a ver a diferença?*). Em PE, este tipo de interrogativa, geralmente, partilha com a declarativa a mesma ordem sintáctica, contrariamente ao inglês que, frequentemente, codifica as interrogativas de sim-não com inversão verbo auxiliar/sujeito (*e.g.*, *Can you see the difference?*). Uma interrogativa alternativa implica duas ou mais hipóteses expressas pela conjunção disjuntiva "ou"

(*e.g.*, *Acha que vai facilitar ou vai ainda tornar mais difícil?*). Uma interrogativa Qu-, por seu turno, é expressa por pronomes e advérbios interrogativos, tais como: *qual*, *quem*, *quando*, *onde*, etc., que correspondem ao constituinte sobre o qual é feita a pergunta (*e.g.*, *Qual é a pergunta?*). Por último, a estrutura de uma interrogativa *tag* corresponde à retoma de uma frase declarativa produzida (*e.g.*, *Isto é fácil, não é?*).

A diversidade de estruturas interrogativas pode implicar a detecção com maior sucesso de determinados tipos em detrimento de outros. O pressuposto de que a estrutura linguística dos diferentes tipos de interrogativas pode, de facto, condicionar o seu processamento automático só recentemente tem sido verificado. Na literatura crítica sobre a detecção de fronteiras de frase, em geral, e sobre a detecção de interrogativas, em particular, tem sido discutido o peso relativo de diferentes tipos de pistas linguísticas para o caso específico do inglês. Wang & Narayanan (2004) defendem que as pistas prosódicas *per se* são as mais eficazes. Por seu turno, Shriberg *et al.* (2009) demonstram que as pistas prosódicas são, de facto, mais significativas do que as lexicais, porém ambas combinadas promovem os melhores resultados. Já para Boakye *et al.* (2009) as pistas léxico-sintácticas são as mais significativas.

Face às opiniões divergentes, questiona-se, no presente trabalho, se o peso relativo das diferentes pistas está directamente relacionado com a natureza dos *corpora* analisados, nomeadamente, se depende dos tipos mais frequentes e específicos de interrogativas nesses *corpora* e das formas como nas línguas são codificados os tipos de frases (relembre-se que, por exemplo, uma interrogativa de sim-não é muitas vezes expressa com mais informação lexical no inglês do que no PE). Para o efeito, foram analisados *corpora* de diálogos espontâneos, de aulas universitárias, de noticiários televisivos e, por uma questão de comparação, de textos jornalísticos.

O artigo está estruturado da seguinte forma: na secção 2 são apresentados estudos linguísticos anteriores sobre as propriedades prosódicas das interrogativas em PE; na secção 3 são descritos os *corpora* analisados; na secção 4 é realizada a análise estatística dos diferentes tipos de interrogativas; as experiências de pontuação automática baseadas em propriedades lexicais e prosódicas são apresentadas na secção 5, e, finalmente, as conclusões e trabalho futuro na secção 6.

## 2. Análise prosódica das interrogativas em PE

As frases declarativas são o tipo de frase mais estudado em PE (Martins, 1986; Viana, 1987; Falé, 1995; Vigário, 1995; Cruz-Ferreira, 1998; Mata, 1999; Frota, 2000; Viana *et al.*, 2007). Os contornos descritos ao longo do presente trabalho estão de acordo com o sistema de anotação Tones and Break Indices (ToBI), proposto e descrito

em Silverman *et al.* (1992); Sun-ah *et al.* (2005); *inter alia*, adaptado para o português por Viana *et al.* (2007) e Frota (2000; 2009). Embora as designações nos diferentes estudos realizados para o português sejam distintas, é consensual que o contorno tipicamente associado a uma declarativa é um contorno descendente. No quadro das adaptações feitas ao sistema ToBI para o português, esse contorno entoacional descendente é expresso por um acento pré-nuclear H* (na primeira sílaba acentuada), um acento nuclear H+L* e um tom de fronteira L% (Frota, 2000). Um contorno similar é reportado para as interrogativas Qu- (Viana, 1987; Frota 2000; Frota, 2009). De acordo com Viana (1987) e Falé (2006) as interrogativas Qu- distinguem-se de uma declarativa por apresentarem uma gama de variação de $f_0$ superior. Ainda sobre as interrogativas Qu-, para Cruz-Ferreira (1998) estas podem também apresentar um contorno ascendente quando enunciadas de forma cortês.

Contrastivamente, o contorno associado a uma interrogativa de sim-não é tipicamente um contorno ascendente, expresso ou como H* H+L* H% ou ainda (H) H+L* LH% (o último contorno foi proposto por Frota, 2002). Mata (1990) observa, por seu turno, contornos descendentes em questões sim-não, mas no contexto específico de fala espontânea.

Tanto quanto nos é possível afirmar, as interrogativas alternativas apenas foram analisadas por Viana (1987) e por Mata (1990). No primeiro estudo, a primeira unidade tonal é representada com um contorno ascendente-descendente-ascendente, sendo o segundo expresso por um contorno ascendente-descendente. No segundo estudo, a primeira unidade corresponde a um contorno baixo-ascendente e a segunda a um baixo--descendente.

Relativamente às interrogativas *tags*, a sua análise prosódica é escassa, apenas Mata (1990), Cruz-Ferreira (1998) e Frota (2000) as analisaram. Mata (1990) descreve-as com tons baixo-ascendentes, enquanto Cruz-Ferreira as caracteriza com tons baixo--descendentes. Frota (2009) descreve a estratégia de fraseamento das *tags* como domínios pós-nucleares acentuados, com registo de $f_0$ comprimido.

Dada a escassez de dados sobre as *tags*, procedeu-se, no âmbito do presente trabalho, a uma análise mais detalhada dessas estruturas nos *corpora*. As interrogativas *tag* têm um comportamento idiossincrático, tanto a nível lexical, quanto a nível prosódico. Distinguem-se lexicalmente por comportarem diferentes expressões com valores pragmáticos distintos (*e.g.*, incerteza, dúvida, surpresa, ou mesmo irritação, *inter alia*), como: *não é?, não é verdade?*, *pois não?*, *certo?*, *correcto?*, *okay?*, *humhum?*, *é isso?*, etc.. Foneticamente, são produzidas com formas fortes e fracas (caracterizadas por uma forte redução vocálica) e podem ser entoacionalmente acentuadas ou não acentuadas.

A análise de uma das *tags* mais frequentes nos *corpora* (*não é?*) evidenciou padrões regulares na sua produção. A parte declarativa apresenta o comportamento expectável de uma declarativa neutra em PE, ou seja, com o contorno H+L* L- ou L%. A *tag* é maioritariamente acentuada e exibe padrões entoacionais diversos, sendo o mais frequente L*+H H%. Quando as *tags* são desacentuadas, apresentam tons de fronteira L% ou LH%. Foneticamente, são maioritariamente produzidas com formas fracas ([nÉ][1] ou [n6 É] *vs.* [n6~w~ É]) e a escolha da forma depende do falante. Acrescente-se que a *tag* é subordinada relativamente ao contorno da declarativa (frequentemente com compressão da frequência fundamental).

## 3. Corpora

De modo a verificar se as propriedades linguísticas são dependentes da natureza do *corpus*, analisaram-se quatro *corpora* distintos. O *corpus* Lectra (Trancoso *et al.*, 2008) foi recolhido no âmbito do projecto homónimo, com o objectivo de transcrever aulas universitárias para aplicações de ensino via *internet*, em especial, para alunos com deficiência auditiva. Integra aulas de 7 disciplinas (6 falantes masculinos e um feminino) e tem aproximadamente 75h, sendo que 27h foram já manualmente anotadas, perfazendo um total de 155 milhares de palavras.

O *corpus* Coral (Viana *et al.*, 1998) compreende 64 diálogos de resolução de uma tarefa, designadamente, a reconstituição de percursos em mapas. Os diálogos foram realizados por 32 falantes, totalizam 7h de fala e compreendem 61 milhares de palavras.

O Alert (Neto *et al.*, 2003) é um *corpus* de notícias televisivas recolhido entre 2000 e 2001. O subconjunto de dados utilizado no presente trabalho compreende 61h e 449 milhares de palavras.

Por uma questão de comparação, também foi utilizado um *corpus* com textos jornalísticos retirados do jornal Público, com 148 milhões de palavras.

Todos os *corpora* foram subdivididos em conjuntos de treino, teste e desenvolvimento, para realização das diferentes experiências. O conjunto de desenvolvimento é importante para afinar o sistema e o conjunto de teste, sendo distinto dos outros dois conjuntos, permite uma avaliação não influenciada pelos dados de treino.

[1] Transcrição fonética representada em alfabeto SAMPA (http://www.phon.ucl.ac.uk/home/sampa/)

## 4. Análise estatística das interrogativas

O quadro 1 mostra a frequência de interrogativas e de outras marcas de pontuação nos diferentes conjuntos de treino dos *corpora*. A frequência absoluta de interrogativas nos dados é substancialmente diferente. Por um lado, as aulas universitárias e os diálogos do Coral compreendem 20,7% e 23,2% de interrogativas, respectivamente; por outro, as notícias televisivas e os jornais apresentam apenas 2,1% e 1,0%, respectivamente. Os dois primeiros *corpora* têm dez vezes mais interrogativas do que os dois últimos, percentagens interpretáveis pela necessidade de o professor verificar, com alguma regularidade, se os alunos estão a perceber a exposição, ou de um dador fornecer as pistas para a reconstituição de um percurso num mapa, procurando clarificar a localização específica.

| ***Corpora*** | **Tipo** | **?** | **!** | **.** | **,** | **:** | **;** | **Frases** |
|---|---|---|---|---|---|---|---|---|
| **Lectra** | Aulas universitárias | 20,7% | 0% | 41,6% | 37,6% | 0% | 0,1% | 6.524 |
| **Coral** | Diálogos espontâneos | 23,2% | 0,4% | 66,9% | 8,0% | 0% | 1,4% | 8.135 |
| **Alert** | Notícias televisivas | 2,1% | 0,1% | 58,1% | 39,1% | 0,5% | 0,2% | 26.467 |
| **Público** | Jornal | 1,0% | 0,2% | 30,7% | 57,5% | 2,4% | 0,7% | 5.841.273 |

**Quadro 1**. Percentagens dos diferentes tipos de pontuação nos conjuntos de treino dos *corpora*.

### 4.1. Frequência de tipos de interrogativas nos *corpora*

A etiquetagem automática dos tipos de interrogativas foi realizada para todos os *corpora* em função do seguinte conjunto de regras heurísticas:

i) se a interrogativa compreende algum dos seguintes itens: *quem*, *qual*, *quais*, *quanto(s)*, *quanta(s)*, *quando*, *quê*, *a quem*, *o quê*, *por que*, *para que*, *onde*, *porque*, *porquê*, *o que*, *como*, então a interrogativa é classificada de Qu-;

ii) se a interrogativa tem uma conjunção coordenada disjuntiva *ou*; então é uma interrogativa alternativa;

iii) se a interrogativa tem um dos seguintes itens: *não é*, *certo*, *não*, *sim*, *okay*, *humhum*, *está bom*, *está*, *está bem*, *não foi*, *tens*, *estás a ver*, *estão a ver*, *é isso*, *de acordo*, *percebo*, *perceberam*, *correcto*, *não é verdade* imediatamente antes de um ponto de interrogação, então é classificada de *tag*,

iv) caso não se verifiquem nenhuma das alíneas anteriores, então a interrogativa é classificada de sim-não.

Com este conjunto de regras heurísticas pretendia-se modelar as expressões lexicais que dão conta dos diferentes tipos de interrogativas. As expressões foram seleccionadas através de uma abordagem baseada nos vastos conjuntos de dados de treino (*data-driven*

*approach*), o que permitiu a inclusão de expressões que não estão ainda devidamente exploradas no PE, sobretudo no que diz respeito às *tag*s, com exemplos como: *okay?* ou *humhum?*.

No quadro 2, encontram-se as frequências absolutas de todos os tipos de interrogativas nos conjuntos de treino dos *corpora* em análise.

| ***Corpora*** | **Tipo** | **Qu-** | **Alt** | ***Tag*s** | **S/N** | **Total de Frases** |
|---|---|---|---|---|---|---|
| **Lectra** | Aulas universitárias | 42,2% | 2,2% | 27,0% | 28,6% | 6.524 |
| **Coral** | Diálogos espontâneos | 7,5% | 3,1% | 12,3% | 77,1% | 8.135 |
| **Alert** | Notícias televisivas | 34,2% | 5,5% | 10,9% | 49,4% | 26.467 |
| **Público** | Jornal | 41,3% | 7,8% | 1,1% | 49,8% | 5.841.273 |

**Quadro 2**. Percentagens dos diferentes tipos de interrogativas nos conjuntos de treino dos *corpora.*

O quadro ilustra que também os diferentes tipos de interrogativas se distribuem de formas distintas. Assim, as aulas universitárias, os noticiários televisivos e os jornais apresentam percentagens similares de interrogativas Qu-. No que concerne às *tag*s, as aulas universitárias são o domínio em que estas são mais significativas. A frequência superior de *tag*s nas aulas universitárias pode explicar-se pelo facto de o professor recorrer a essa estrutura para confirmar se os alunos estão a compreender os conteúdos leccionados. No *corpus* de diálogos espontâneos as *tag*s também são representativas, porém as interrogativas sim-não suplantam, em larga medida, todos os outros tipos. Esta estratégia prende-se com a necessidade de um dador questionar o seu seguidor sobre o local específico do mapa em que este se encontra, pretendendo uma resposta afirmativa ou negativa, para prosseguir o diálogo e consequentemente solucionar o percurso. A frequência dos tipos de interrogativas nas notícias televisivas é bastante similar à dos telejornais, tal como expectável. Quanto às interrogativas alternativas, a sua produção é residual nos *corpora* em análise.

### 4.2. Colecções douradas dos *corpora*

De modo a criar um conjunto de referência para cada *corpus* com a classificação de uma amostra satisfatória de diferentes tipos de interrogativas, foram construídas colecções douradas. O processo de construção das referidas colecções partiu da etiquetagem automática e da sua posterior correcção. A escolha deste processo está associada ao facto de, em simultâneo, se corrigir os erros provenientes da classificação automática e, desta forma, construir uma base de dados para teste.

No Quadro 3 são apresentados os resultados da etiquetagem automática e da correcção manual para os subconjuntos de teste. A concordância entre os resultados obtidos em ambas as partes do processo foi medida com o coeficiente Cohen's Kappa (Carletta, 1996), valores apresentados na última coluna do quadro. Discriminando os valores de concordância por tipo de interrogativa, verifica-se que as interrogativas alternativas são mais bem classificadas, com um coeficiente de 0,912, seguidas das interrogativas Qu- (0,874) e das de sim-não, com resultados similares (0,863). Tal como esperado, as inconsistências mais notórias dizem respeito à classificação das *tag*s (0,782).

| | | | **Classificação Automática** | | | | **Classificação Manual** | | | | **Cohen's** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **#Frase** | **#?** | **Qu** | **Alt** | ***Tag*** | **Tot** | **Qu** | **Alt** | ***Tag*** | **Tot** | **Kappa** |
| **Lectra** | 262 | 102 | 39,7 | 2,2 | 39,0 | 19,1 | 41,4 | 1,0 | 40,4 | 17,1 | 0,922 |
| **Coral** | 3.406 | 511 | 9,4 | 3,5 | 13,5 | 73,6 | 10,6 | 5,1 | 18,2 | 66,1 | 0,849 |
| **Alert** | 2.671 | 151 | 42,4 | 2,0 | 11,2 | 44,4 | 40,4 | 2,6 | 10,0 | 47,0 | 0,895 |
| **Púb** | 90.534 | 2.859 | 44,9 | 7,3 | 0,9 | 46,9 | 43,5 | 6,3 | 0,8 | 49,4 | 0,900 |

**Quadro 3**. Classificações automática e manual dos subconjuntos de teste dos *corpora*.

Como se verifica no Quadro 3, as regras foram aplicadas de um modo bastante satisfatório, com valores do coeficiente Cohen's Kappa que rondam os 90%. A obtenção de níveis de concordância de 85% no *corpus* de diálogos está intrinsecamente associada a estruturas que tanto podem ser classificadas de interrogativas sim-não elípticas (*e.g.*, sim?, é?) ou de *tag*s (*e.g.*, declarativa + sim?, declarativa + é?), dependendo a sua diferenciação de uma análise mais abrangente do contexto morfossintáctico em que as mesmas estruturas se inserem. O *corpus* de aulas universitárias, por seu turno, é o que apresenta valores mais elevados de concordância (92,2%), devendo-se os erros a classificações erróneas similares às do *corpus* de diálogos. As notícias televisivas e os textos jornalísticos contêm estruturas mais complexas e, consequentemente, mais difíceis de desambiguar automaticamente (*e.g.*, orações complexas encaixadas), e apresentam valores de concordância muito próximos, demonstrando que também nesse ponto se assemelham.

Desta forma, poder-se-á concluir que, por um lado, tal como era esperado, as notícias televisivas e os textos jornalísticos são mais similares no que diz respeito à frequência tanto dos diferentes tipos de interrogativas, em particular, quanto das interrogativas, em geral, bem como da própria natureza dos erros; por outro, as aulas universitárias e os diálogos partilham estruturas mais flexíveis, próprias da fala espontânea, tais como as interrogativas *tag*s e as de sim-não elípticas.

Dos resultados apresentados, poder-se-á concluir pela presença de duas estratégias principais nos *corpora*: a da apresentação que não depende de um interlocutor e a da interacção vívida, centrada na clarificação recorrente (*e.g.*, através de *tag*s) e na cortesia de auxílio ao interlocutor, quando este poderá estar a não encontrar o local exacto num mapa (*e.g.*, interrogativas de sim-não).

## 5. Tarefa de pontuação automática

Esta secção diz respeito à tarefa de detecção automática de interrogativas com recurso a diversas pistas linguísticas. Primeiramente, dar-se-á conta do impacto das pistas lexicais, aprendidas através de um vasto *corpus* de textos jornalísticos; posteriormente, far-se-á a análise do contributo de cada pista prosódica discriminadamente.

O módulo de pontuação automática foi inicialmente concebido (Batista *et al.*, 2008) para detectar apenas os pontos finais e as vírgulas. Com o presente trabalho pretende-se alargar a tarefa de pontuação para que possa abranger também o ponto de interrogação. O módulo em questão baseia-se em modelos de Máxima Entropia (ME), cuja aplicação permite explorar e combinar diferentes propriedades da informação a tratar para um determinado tópico. Este método é de especial interesse para esta tarefa, visto tirar partido do vasto conjunto de propriedades linguísticas disponíveis, quer lexicais, quer prosódicas. A ferramenta MegaM (Daumë, 2004), de domínio público, foi utilizada para treinar os modelos com base na abordagem ME. Os resultados foram avaliados de acordo com as métricas-padrão, designadamente, as de precisão (o número de pontos de interrogação correctos sobre o número de pontos de interrogação na hipótese), cobertura (o número de pontos de interrogação correctos sobre o número total de pontos de interrogação na referência), medida-F= $\frac{2 \text{ x a precisão x a cobertura}}{\text{a precisão x a cobertura}}$ e *Slot Error Rate* (SER, corresponde a uma taxa de erro, mais concretamente ao número de erros na identificação das interrogativas dividido pelo número de interrogativas existentes na referência).

A anotação manual é realizada ao nível da frase, sendo necessário ajustar para os níveis da palavra e dos fones. Este processo é feito automaticamente com recurso ao módulo de Reconhecimento Automático de Fala (Amaral *et al*., 2007) no modo de alinhamento forçado.

Ainda assim, nem sempre é possível o alinhamento de todas as palavras, devido a distintos factores, como por exemplo, partes do sinal com energia insuficiente, ou zonas

de fala sobreposta. Por essa razão, o conjunto de teste do *corpus* de notícias televisivas tem 1% de erro no alinhamento forçado, sendo o erro no *corpus* de aulas universitárias superior (5,3%). O *corpus* de diálogos espontâneos não foi utilizado na tarefa de pontuação devido à significativa percentagem de fala sobreposta, cujo alinhamento com o sinal não se encontra demarcado manualmente[2]. A pontuação tida como referência para ambos os *corpora* de notícias televisivas e aulas universitárias corresponde à pontuação dos dados anotados manualmente. Para a obtenção da pontuação de referência foi utilizada a ferramenta NIST SCLite[3], com uma etapa de pós--processamento, para correcção de erros posteriormente detectados.

### 5.1. Pistas lexicais

De forma a poder comparar as diferentes experiências, foi criado um modelo inicial, utilizando apenas a informação dos textos jornalísticos, *i.e.*, 143 milhões de palavras. Para cada uma das frases do corpus foram utilizadas as seguintes propriedades: cada uma das palavras da frase; bigramas e trigramas de palavras adjacentes; informação sobre as palavras, bigramas e trigramas encontradas no início e no fim da frase e o número de palavras da frase. Os resultados obtidos neste processo estão contemplados no Quadro 4. Os valores na coluna dos "Correctos" correspondem a marcas de pontuação correctamente identificadas, por seu turno, a dos "Incorrectos" compreende, simultaneamente, falsos positivos ou inserções; a das "Omissões" dá conta de apagamentos ou de não identificação de marcas de pontuação.

| *Corpora* | # Frases | Cor | Inc | Omi | Prec | Cob | F | SER |
|---|---|---|---|---|---|---|---|---|
| **Lectra** | 1.120 | 158 | 32 | 220 | 83,2% | 41,8% | 55,6% | 66,7% |
| **Alert** | 9.552 | 128 | 25 | 287 | 83,7% | 30,8% | 45,1% | 75,2% |
| **Público** | 222.127 | 1100 | 236 | 1740 | 82,3% | 38,7% | 52,7% | 69,6% |

**Quadro 4**. Resultados obtidos exclusivamente a partir de pistas lexicais. "Cor" corresponde a Correctos, "Inc" a Incorrectos, "Omi" a omissões, "Prec" a precisão, "Cob" a cobertura, "F" a medida-F e "SER" a *slot error rate*.

Como se pode verificar no Quadro 4, a precisão de todos os *corpora* é aproximadamente de 83%, porém a cobertura é assaz inferior. A conclusão a retirar dos resultados obtidos na criação da base de referência, conseguida exclusivamente a partir de pistas lexicais, é a de que a percentagem de cobertura está correlacionada com a

[2] No corpus Coral, a tarefa de demarcação manual da fala sobreposta, devidamente alinhada com o sinal, só recentemente foi concluída.

[3] http://www.itl.nist.gov

identificação de um tipo de interrogativa específica, as Qu-. Neste ponto, é importante recordar que a percentagem de cobertura é comparável à da distribuição de interrogativas Qu- nos diferentes *corpora*. Acrescente-se que, tendo em conta apenas as pistas lexicais, as interrogativas sim-não só residualmente são identificadas.

### 5.2. Pistas de segmentação áudio

Numa segunda etapa, as transcrições efectivas de cada *corpus*, obtidas por alinhamento forçado, foram usadas para retreinar o modelo inicial, explicitado na secção anterior. Com este alinhamento forçado, foi possível adicionar às pistas lexicais, descritas na secção anterior, informação sobre a mudança de falante e sobre a duração dos silêncios.

O Quadro 5 mostra os resultados da inclusão desstas pistas. O dado mais significativo na leitura do quadro prende-se com a melhoria performativa em ambos os *corpora*, especialmente para o *corpus* de aulas universitárias, uma vez que neste a pista mudança de falante não é tão produtiva, sendo a duração dos silêncios muito mais informativa para a identificação de marcas de pontuação.

| ***Corpora*** | **Cor** | **Inc** | **Omi** | **Prec** | **Cob** | **F** | **SER** |
|---|---|---|---|---|---|---|---|
| **Lectra** | 271 | 52 | 107 | 83,9% | 71,7% | 77,3% | 42,1% |
| **Alert** | 144 | 27 | 271 | 84,2% | 34,7% | 49,1% | 71,8% |

**Quadro 5**. Resultados obtidos após o retreino com as transcrições e com o acréscimo de pistas relativas ao falante.

### 5.3. Pistas prosódicas

Numa terceira etapa, procurou-se analisar o peso relativo e o contributo de cada propriedade prosódica *per se* e aferir do impacto da combinação de todas as pistas prosódicas. Esta tarefa assenta em evidências linguísticas de que os contornos nucleares, os tons de fronteira, os declives de energia, a presença/ausência de pausas e a sua duração são essenciais para a delimitação de frases e a caracterização dos tipos de frases em várias línguas.

Primeiramente, testaram-se as unidades de análise a ter em conta e a importância das mesmas em função das distintas pistas prosódicas. Desta forma, verificou-se se os valores acústicos a extrair eram mais relevantes ao nível do fone, da sílaba ou da palavra. Com base em evidências do PE, colocou-se a hipótese de que as unidades de análise mais relevantes seriam as sílabas acentuadas e pós-acentuadas antes de um silêncio ou de uma ruptura melódica.

A extracção dos parâmetros acústicos envolveu uma série de etapas. Numa primeira fase, a frequência fundamental e a energia foram extraídas do sinal com recurso à ferramenta de domínio público Snack Sound Toolkit[4]. As durações dos fones, das palavras e dos intervalos entre palavras foram obtidas através do reconhecedor. Uma vez extraídos os valores de frequência fundamental e as fronteiras dos fones, efectuou-se uma etapa de pós-processamento, removendo saltos de oitava e efeitos de micro-prosódia. Dado que o reconhecedor não possui um módulo de silabificação e de acentuação, foi necessário criar um conjunto de regras de silabificação e de acentuação a aplicar ao léxico. As regras construídas, no âmbito do presente trabalho, têm um resultado performativo bastante satisfatório para palavras nativas, porém necessitam de ajustes para dar conta de estrangeirismos. Finalmente foram calculados os valores máximos, mínimos, medianas e declives de $f_0$ (normalização numa escala de semitons) e de energia para cada unidade de análise (fone, sílaba e palavra).

As pistas são calculadas para cada transição entre frases, com ou sem pausa, usando o mesmo escopo de análise de Shriberg *et al.* (2009), *i.e.*, última palavra, última sílaba acentuada e último fone vozeado de uma dada fronteira e primeira palavra, primeiro fone vozeado da unidade seguinte. Deu-se especial relevo às pistas: declives de energia e de $f_0$ numa palavra antes e depois de uma pausa, diferenças de energia e de $f_0$ entre as palavras em questão e a duração da última sílaba e do último fone. O peso dado a este subconjunto de propriedades linguísticas está intrinsecamente associado à modelação de contornos nucleares, tons de fronteira e de potencial alongamento final. As pistas prosódicas descritas foram previamente aplicadas à tarefa de identificação de pontos finais e vírgulas com resultados significativos (melhorias de 2% absolutos) para o *corpus* de notícias televisivas (Batista *et al.*, 2010).

Os resultados da identificação de interrogativas com recurso a pistas prosódicas nos *corpora* de aulas universitárias e de notícias televisivas são apresentados nos quadros 6 e 7, respectivamente. Quando se comparam os resultados obtidos com a introdução das propriedades prosódicas, verifica-se que houve uma melhoria, sobretudo no *corpus* de aulas universitárias.

[4] http://www.speech.kth.se/snack/

| Unidade | Pistas | Cor | Inc | Omi | Prec | Cob | F | SER |
|---|---|---|---|---|---|---|---|---|
| **Palavra** | $f_0$ | 275 | 53 | 103 | 83,8% | 72,8% | 77,9% | 41,3% |
| | energia | 266 | 54 | 112 | 83,1% | 70,4% | 76,2% | 43,9% |
| | $f_0$ e energia | 273 | 52 | 105 | 84,0% | 72,2% | 77,7% | 41,5% |
| **Sílabas e fones** | $f_0$ | 269 | 54 | 109 | 83,3% | 71,2% | 76,7% | 43,1% |
| | energia | 269 | 49 | 109 | 84,6% | 71,2% | 77,3% | 41,8% |
| | duração | 268 | 52 | 110 | 83,8% | 70,9% | 76,8% | 42,9% |
| | $f_0$, energia e duração | 268 | 50 | 110 | 84,3% | 70,9% | 77,0% | 42,3% |
| **Todas** | | 273 | 50 | 105 | 84,5% | 72,2% | 77,9% | 41,0% |

**Quadro 6**. Identificação de interrogativas com recurso a informação prosódica no *corpus* de aulas universitárias.

| Unidade | Pistas | Cor | Inc | Omi | Prec | Cob | F | SER |
|---|---|---|---|---|---|---|---|---|
| **Palavra** | $f_0$ | 149 | 27 | 266 | 84,7% | 35,9% | 50,4% | 70,6% |
| | energia | 146 | 25 | 269 | 85,4% | 35,2% | 49,8% | 70,8% |
| | $f_0$ e energia | 147 | 27 | 268 | 84,5% | 35,4% | 49,9% | 71,1% |
| **Sílabas e fones** | $f_0$ | 151 | 27 | 264 | 84,8% | 36,4% | 50,9% | 70,1% |
| | energia | 146 | 24 | 269 | 85,9% | 35,2% | 49,9% | 70,6% |
| | duração | 144 | 28 | 271 | 83,7% | 34,7% | 49,1% | 72,0% |
| | $f_0$, energia e duração | 147 | 29 | 268 | 83,5% | 35,4% | 49,7% | 71,6% |
| **Todas** | | 146 | 28 | 269 | 83,9% | 35,2% | 49,6% | 71,6% |

**Quadro 7**. Identificação de interrogativas com recurso a informação prosódica no *corpus* de notícias televisivas.

Os resultados são parcialmente concordantes com os de Shriberg *et al.* (2009) no que diz respeito ao contributo de cada pista, sendo o declive de $f_0$ da palavra antes de um silêncio e a relação desse mesmo declive com o da palavra seguinte as pistas mais significativas. Tal como reportado por Vassière (1983), essas pistas são independentes da língua. Dados especificamente relacionados com o PE são, sobretudo, os padrões duracionais, as realizações fonéticas de determinados fones, como as fricatizações de oclusivas nos referidos contextos (Moniz *et al.*, 2010), bem como os declives de $f_0$ associados a funções discursivas distintas, que estão para além da simples forma das frases (para as interrogativas confirmativas, por exemplo, veja-se Mata & Santos, 2010).

Em jeito de conclusão, quando se treina apenas com pistas lexicais, as interrogativas Qu- são identificadas com algum relevo, enquanto as *tag*s e as sim-não só

residualmente o são, exceptuando a sequência "acha que". Contudo, persistem omissões na identificação de interrogativas Qu-, devido a estruturas complexas cuja desambiguação automática é ainda desafiante. Quando são adicionadas as pistas prosódicas, então as interrogativas sim-não passam a ser identificadas. Acrescente-se que as pistas prosódicas contribuem positivamente para a identificação de interrogativas em ambos os *corpora*, mesmo quando as interrogativas correspondem a estruturas sincopadas, como o pedido para repetir informação (*e.g.*, "recta das?"), ou a interrogativas *tag*s (*e.g.*, "certo?").

Deve, contudo, referir-se que, embora os resultados obtidos sejam positivos, subsistem erros de precisão e de cobertura, devido a um conjunto de estruturas que colocam desafios ao processamento automático de fala, a saber:

i) um número considerável de questões realizadas na transição entre *pivot* e repórter, com condições de fundo ruidosas (*e.g.*, cenários bélicos);

ii) questões elípticas frequentes (*e.g.*, *Eu?*; *Sim?* );

iii) sequências com disfluências (*e.g.*, questão com disfluências não identificada, *<é é é> como é que consegue? vs.* Questão sem disfluências devidamente detectada, *Como é que conseguem isso?*);

iv) falsos positivos com interrogativas subordinadas que não são pontuadas com ponto de interrogação (*e.g.*, *perguntou se vinhas hoje.*);

v) sequências com mais do que uma interrogativa consecutiva (*e.g.*, *nascem duas perguntas: quem? E porquê?*);

vi) sequências que integram parentéticas ou vocativos (*e.g.*, *Foi acidente mesmo ou atentado, Noé?*).

## 6. Conclusões e trabalho futuro

As hipóteses inicialmente colocadas de que a distribuição das interrogativas, em geral, e dos seus subtipos, em particular, dependiam da natureza dos *corpora* foram confirmadas, como atestam os resultados apresentados anteriormente.

O conjunto de regras especificamente criado para identificar automaticamente os diferentes tipos de interrogativas dá conta das distinções lexicais entre elas de forma bastante satisfatória. Estas regras serviram, sobretudo, o propósito de criar colecções douradas para testar o processamento posterior das interrogativas.

O processo de identificação automática de interrogativas foi faseado, de modo a aferir o contributo das diversas pistas linguísticas. Primeiramente, foram treinados modelos com base em informação lexical proveniente de um extenso *corpus* de textos jornalísticos. Os resultados desta fase inicial demonstraram que apenas as interrogativas

QU- são expressivamente identificadas, com percentagens de identificação similares à distribuição das referidas interrogativas nos diferentes *corpora*. Posteriormente, foram treinados modelos com base nas transcrições efectivas dos *corpora*, com recurso a pistas lexicais e a pistas relativas ao falante. Os resultados obtidos foram melhores, porém determinados tipos de interrogativas (*tag*s e sim-não) continuaram a não ser identificados. Finalmente, quando foram treinados modelos com base em pistas prosódicas, então os referidos tipos, até então omissos, passaram a ser identificados. Os resultados mais expressivos são conseguidos com a combinação de todas as pistas linguísticas empregues nas diferentes etapas.

Ainda que o processamento automático de interrogativas esteja no seu preâmbulo, deve salientar-se que, tanto quanto nos é possível afirmar, o presente estudo é uma primeira quantificação de tipos de interrogativas em diversos *corpora* e uma primeira discussão do contributo de distintas pistas linguísticas na detecção de interrogativas em fala preparada e espontânea em PE.

Em trabalho futuro, pretende-se alargar a análise a outras línguas, nomeadamente, ao inglês e ao castelhano, de modo a aferir da (in)dependência dos resultados.

## Referências


Amaral, R., H. Meinedo, D. Caseiro, I. Trancoso & J. Neto (2007) A prototype system for selective dissemination of broadcast news in European Portuguese. In *EURASIP, Journal of Advances in Signal Processing*. Hindawi Publishing Corporation, vol. 2007, n. 37507.

Batista, F., D. Caseiro, N. Mamede & I. Trancoso (2008) Recovering capitalization and punctuation marks for automatic speech recognition: case study for Portuguese broadcast news. In *Speech Communication*, vol. 50, pp. 847-862.

Batista, F., H. Moniz, I. Trancoso, H. Meinedo, A. I. Mata & N. Mamede (2010) Extending the punctuation module for European Portuguese. In *Interspeech 2010*, Makuhari, Japão.

Beckman, M. *et al.* (2005) The Original ToBI System and the Evolution of the ToBI Framework. In S. Jun (ed.) *Prosodic Typology. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

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

Carletta, J. (1996) Assessing agreement on classification tasks: the kappa statistics. In *Computational Linguistics*, 22(2), pp. 249-254.

Cruz-Ferreira, M. (1998) Intonation in European Portuguese. In D. Hirst & A. Di Cristo (eds.) *Intonation systems – a survey of twenty languages*, Cambridge University Press.

Daumé III, H. (2004) Notes on CG and LM-BFGS optimization of logistic regression. http://hal3.name/megam/

Falé, I. (1995) *Fragmento da prosódia do português europeu: as estruturas coordenadas*. Tese de Mestrado, FLUL.

Falé, I. (2006) *Percepção e Reconhecimento da Informação Entoacional em Português Europeu*. Dissertação de Doutoramento, Universidade de Lisboa.

Frota, S. (2000) *Prosody and Focus in European Portuguese*. New York & London: Garland Publishing.

Frota, S. (2002) Nuclear falls and rises in European Portuguese. In *Probus*, pp. 113-146.

Frota, S. (2009) The intonational phonology of European Portuguese. In S. Jun (ed.) *Prosodic Typology II: The Phonology and Intonation of Phrasing*. Oxford: Oxford University Press.

Hirschberg, J. & J. Pierrehumbert (1986) The intonational structuring of discourse. In *The 24th Annual Meeting of the Association for Computational Linguistics.* New York: Columbia University.

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

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

Martins, F. (1986) *Entoação e organização do enunciado*. Tese de Mestrado, FLUL.

Mateus, M. *et al.* (2003) *Gramática da Língua Portuguesa*. Lisboa: Caminho.

Mata, A. I. (1990) *Questões de entoação e interrogação no português*. Tese de Mestrado, FLUL.

Mata, A. I. (1999) *Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologias, Resultados e Implicações Didácticas*. Dissertação de Doutoramento, FLUL.

Mata, A. I. & A. Santos (2010) On the intonation of confirmation-seeking requests in child-directed speech. In *Speech Prosody 2010*, Chicago.

Moniz, H., F. Batista, H. Meinedo, A. Abad, I. Trancoso & A. I. Mata (2010) Prosodically-based automatic segmentation and punctuation. In *Speech Prosody 2010*, Chicago.

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

Shriberg, E. (1998) Can prosody aid the automatic classification of dialog acts in conversational speech?. In *Language and Speech*, pp. 439-487.

Shriberg, E., B. Favre, J. Fung, D. Hakkani-Tür, S. & Cuendet (2009). Prosodic similarities of dialog act boundaries across speaking styles. In S. C. Tseng (ed.), *Linguistic Patterns in Spontaneous Speech*, Taipei: Institute of Linguistics, Academia Sinica, pp. 213-239.

Silverman, K., M. Beckam, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert & J. Hirschberg (1992) ToBI: a standard for labeling English prosody. *In Proceedings ICSLP*, Banff, vol. 2, pp. 867-870.

Sun-Ah Jun (ed.) *Prosodic Typology II*. Oxford: Oxford University Press.

Trancoso, I., R. Martins, H. Moniz, A. I. Mata & M. C. Viana (2008) The Lectra *corpus* – classroom lecture transcriptions in European Portuguese. *In LREC 2008*, Marrocos.

Vassière, J. (1983) Language-independent prosodic features. In A. Cutler (ed.) *Prosody: modules and measurements*. Berlim: Springer, pp. 55-66.

Viana, M. C. (1987) *Para a Síntese da Entoação do Português*. Dissertação de Doutoramento, FLUL.

Viana, M. C., I. Trancoso, I. Mascarenhas, I. Duarte, G. Matos, L. Oliveira, H. Campos & C. Correia (1998) Apresentação do projecto Coral – *corpus* de diálogo etiquetado. In *Workshop de Linguística Computacional.* Lisboa.

Viana, M. C., S. Frota, I. Falé, F. Fernandes, I. Mascarenhas, A. I. Mata, H. Moniz & M. Vigário (2007). Towards a P_ToBI. *PAPI2007. Workshop on the Transcription of*

*Intonation in Ibero-Romance*. Universidade do Minho. (ver http://www.ling.ohio-state.edu/~tobi/).

Vigário, M. (1995) *Aspectos da prosódia do português europeu. Estruturas com advérbios de exclusão e negação frásica*. Tese de Mestrado, Universidade do Minho.

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