Os *corpora* sonoros do grupo da Variação do CLUL

João Saramago
Centro de Linguística da Universidade de Lisboa

Os *corpora* sonoros do grupo da Variação do CLUL já foram apresentados e caracterizados no decorrer do XI Encontro Nacional da APL (Lisboa 1995) por Maria Luísa Segura da Cruz.

Já nessa altura, ela chamava a atenção para o facto de a classificação como *corpus*, do acervo sonoro do grupo da Variação, dever ser entendida no sentido mais amplo e tradicional do termo, isto é, como "um conjunto de dados reunidos com o objectivo de servir determinada pesquisa"[1].

Assim, depois de voltar a fazer uma rápida apresentação de cada um deles, referirei o que de novo existe no âmbito desses *corpora*, sobretudo aquele que constitui o *corpus* do *Atlas-Linguístico-Etnográfico de Portugal e da Galiza (ALEPG)* e a utilização que dele tem sido feita.

Como, regra geral, os *corpora* que são constituídos no âmbito da Geografia Linguística têm como finalidade principal a elaboração de atlas linguísticos de uma determinada área geográfica ou linguística, eles devem obedecer a características específicas. Essas características são essencialmente a homogeneidade que se verifica na sua recolha de modo a permitir dar conta da variação dialectal existente nessas áreas e a uniformidade de critérios na escolha das localidades e dos informantes a inquirir.

A homogeneidade do material recolhido é conseguida através do recurso a um mesmo questionário linguístico cuja aplicação permite, por um lado, a caracterização de cada uma das variedades e, por outro, a confrontação de cada uma das variedades com as restantes.

A uniformidade consegue-se com a escolha criteriosa das localidades que constituem os pontos da rede do Atlas e dos informantes escolhidos em cada uma delas. Relativamente às localidades, a sua escolha prende-se essencialmente com a existência de uma maior ou menor variação linguística e com uma maior ou menor densidade populacional. Quanto aos informantes, a idade, a escolarização, o conhecimento dos assuntos a tratar, a sua capacidade de resposta e o contacto com outras variedades linguísticas, são alguns dos aspectos que condicionam a sua selecção.

Até determinada altura, e mesmo ainda actualmente em alguns casos, os *corpora* que servem de base à realização dos Atlas Linguísticos são constituídos

[1] Maria Luisa Segura da Cruz (1996), 'Os *corpora* dialectais do CLUL: sua caracterização e objectivos', *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística* (Lisboa, 1995), Vol. I, Lisboa, p.151.

***Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2002, pp. 737-742.**

pelos cadernos de inquérito preenchidos localmente pelos inquiridores. Desde há alguns anos que se passou a recorrer à gravação simultânea, integral ou parcial, dos inquéritos. Essas gravações tornam-se amostras vivas das variedades linguísticas e, simultaneamente, vêm permitir outro tipo de estudos, tais como estudos de fonética experimental, de sintaxe e de morfologia, que não seria possível realizar apenas com base em material transcrito.

Começarei por referir os *corpora* do grupo da Variação que foram constituídos para projectos de menor envergadura que o ALEPG e de natureza mais específica.

1. Atlas Linguarum Europae (ALE)

Este é um projecto de âmbito internacional que recobre todo o continente europeu, no qual Portugal participa desde 1974[2].

Até agora já foram publicados cinco fascículos, num total de 59 mapas lexicais acompanhados dos respectivos comentários.

O questionário consta de 546 perguntas de índole lexical e foi aplicado, durante o ano de 1975, em 53 localidades de Portugal continental.

O *corpus* deste projecto atinge cerca de 140 horas de gravação.

2. Atlas linguístico do litoral português (ALLP)

Trata-se de um projecto, da responsabilidade de Gabriela Vitorino, que tem como objectivo o estudo do léxico especializado associado à vida piscatória.

O questionário contém cerca de 1200 perguntas e abrange todos os campos semânticos directamente relacionados com a actividade piscatória.

A rede de pontos é constituída por 40 localidades que têm a pesca como principal actividade: 23 em Portugal continental, 12 nos Açores e 5 na Madeira.

O questionário completo foi já aplicado em 8 localidades. Os materiais referentes à fauna e flora marinhas encontram-se recolhidos na totalidade dos pontos da rede e já deu origem à elaboração de um primeiro volume que abrange os 23 pontos continentais. A obra consta de dois tomos: um, de introdução, dialectometria e índices e outro, de mapas e notas[3].

As recolhas no Continente foram efectuadas em 1986 e 1987, na Madeira em 1994 e nos Açores em 1991, 1995 e 1996.

Presentemente os materiais recolhidos estão a ser introduzidos numa base de dados e prevê-se que os dados referentes aos Açores venham a ser publicados num dos volumes do *Atlas Linguístico-Etnográfico dos Açores (ALEAç).*

O *corpus* deste projecto já atinge as 210 horas de gravação.

---

[2] O ALE foi criado por iniciativa da Universidade Católica de Nijmegen (Holanda), com apoio da UNESCO e conta com a participação de todos os países europeus. Actualmente tem a sua sede em Kiel (Alemanha).

[3] Gabriela Vitorino (1987), *Atlas linguístico do litoral português. Fauna e Flora.* Dissertação apresentada para progressão na carreira de investigação. INIC/CLUL (inédita).

3. Barlavento do Algarve (BA)

O objectivo deste projecto, da responsabilidade de Maria Luísa Segura da Cruz, foi o de estudar o vocalismo de uma variedade dialectal, o Barlavento do Algarve, com a finalidade de traçar as suas actuais fronteiras linguísticas. Este estudo foi completado com uma análise acústica, tendo em vista determinar a estrutura formancial dos fonemas vocálicos[4].

O questionário, que consta de 378 perguntas, foi aplicado total ou parcialmente em 53 localidades do Barlavento algarvio em 1986 e 1987.

O *corpus*, que também contém gravações complementares versando temas como o cultivo da terra, as técnicas e instrumentos de trabalho, ultrapassa as 100 horas de gravação.

4. O falar da ilha do Corvo

Este projecto, da minha responsabilidade, consta do estudo do dialecto corvino sob três perspectivas: um estudo acústico, baseado em análise espectrográfica, das vogais orais acentuadas e não-acentuadas; um estudo lexical tendo em vista a variação interna do dialecto com base na dicotomia léxico dialectal / léxico não-dialectal de acordo com a idade e sexo dos informantes e um estudo dialectométrico que visa relacionar este dialecto com outros dialectos açorianos e com o espaço dialectal continental[5].

O questionário fonético foi aplicado a três informantes masculinos pertencentes aos três escalões etários considerados: 25-30 anos, 45-50 anos e mais de 70 anos. O questionário lexical, com 908 perguntas, foi aplicado a 18 informantes, 9 homens e 9 mulheres, agrupados pelos três escalões etários já referidos (3 informantes masculinos e três femininos por escalão). Os inquéritos foram feitos em finais de 1984 e início de 1985.

O *corpus* atinge as 75 horas de gravação.

5. Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG)

Este projecto teve o seu início em 1970 sob a direcção de Luís F. Lindley Cintra e tem como objectivo principal a publicação do atlas linguístico nacional.

O questionário linguístico, que está organizado por campos semânticos, contém cerca de 4000 perguntas. A rede de pontos definitiva é constituída por 212 localidades assim distribuídas: Portugal continental, 176 localidades; arquipélago da Madeira, 7 localidades, arquipélago dos Açores, 17 localidades e pontos fronteiriços, 12 localidades em território politicamente espanhol. Actualmente falta inquirir 2 pontos em território galego.

---

[4] Maria Luisa Segura da Cruz (1987), *A fronteira dialectal do Barlavento do Algarve*. Dissertação apresentada para progressão na carreira de investigação. INIC/CLUL (inédita).

[5] João Saramago (1992), *Le parler de l'île de Corvo (Açores)*, Centre de Dialectologie, Université Stendhal, Grenoble III, Grenoble.

Em 70 dessas localidades, foi aplicado o questionário integral. No entanto, a partir de 1990, foi decidido reduzir o número de perguntas para cerca de metade, tendo em vista a necessidade de assegurar, o mais rapidamente possível, a cobertura total da rede. Os campos semânticos inquiridos passaram a ser os seguintes: as plantas (ervas, arbustos e flores), os animais (animais domésticos – o cão e o gato, os animais bravios, as aves, os insectos e outros invertebrados, os batráquios e répteis), o homem e o trabalho (a gricultura, o aproveitamento dos produtos vegetais, a criação de gado, ofícios e profissões e outras actividades).

O *corpus* atinge cerca de 3500 horas de gravação que correspondem aos 210 inquéritos realizados entre 1973 e 2000 e que foram gravados na íntegra.

Este *corpus* serve igualmente dois outros projectos, o *Atlas Linguistique Roman (ALiR)* e o *Atlas Linguístico-Etnográfico dos Açores (ALEAç)*.

O *ALiR*, que engloba a totalidade dos países europeus de línguas românicas, é um projecto no qual Portugal participa desde a sua criação em 1987.

O questionário lexical contém 292 perguntas, o questionário de fonética histórica 284 perguntas e o questionário morfosintáctico, 42 perguntas.

A rede portugesa é de 110 pontos: 96 pontos em Portugal continental, 4 na Madeira e 10 nos Açores (estes pontos fazem parte da rede do ALEPG).

Já se encontram publicados dois fascículos de mapas e de comentários num total de 30 mapas.

O *ALEAç* é um projecto, subsidiado pela Direcção Regional da Cultura da Região Autónoma dos Açores, que prevê a publicação, em 8 volumes, do material recolhido nas 17 localidades do arquipélago açoriano que constituem a rede do ALEPG. Como acima fiz referência no projecto *ALLP*, está igualmente prevista a publicação de um nono volume dedicado à fauna e flora marinhas do arquipélago.

O primeiro volume, dedicado à criação do gado, num total de 136 mapas lexicais e de 8 mapas morfofonológicos, encontra-se na tipografia. O segundo volume, que estuda a nomenclatura do vinho e da vinha e dos trabalhos do linho e da lã, num total de 127 mapas lexicais e de 5 morfofonológicos, já se encontra integralmente redigido. Em fase avançada de elaboração, está o terceiro volume dedicado ao cultivo dos cereais, à moagem e à panificação.

6. Outros projectos que estudam os *corpora* sonoros do grupo da Variação

As gravações que compõem os *corpora* acima descritos, para além de compreenderem as respostas obtidas através da aplicação dos questionários linguísticos, compreendem igualmente extensos registos de discurso espontâneo.

Esses excertos de discurso livre constituem um excelente material para a realização de outro tipo de estudos linguísticos, nomeadamente de sintaxe e de morfologia dialectais.

Actualmente existem dois projectos que estudam o material nessa perspectiva.

6.1 *Corpus* dialectal com anotação sintáctica (CORDIAL-SIN)

Este projecto, da responsabilidade de Ana Maria Martins. existe desde 1998, com financiamento da FCT [PRAXIS/P/PLP/33275 e POSI/1999/PLP/33275], e encontra-se na sua segunda fase de elaboração.

O seu objectivo é estudar a variação sintáctica dialectal do Português Europeu, numa perspectiva de Princípios e Parâmetros, com recurso a uma metodologia de constituição/exploração de *corpora*.

Os dados serão disponibilizados em quatro versões:

- transcrição conservadora, que contém informação sobre aspectos da fonte sonora tais como pausas, hesitações, variantes fonéticas e morfológicas, sobreposições de produção, sequências imperceptíveis, etc.;
- transcrição normalizada, obtida a partir da transcrição conservadora através da extracção automática dos códigos que identificam marcas de oralidade (presentemente, quer em transcrição conservadora, quer em transcrição normalizada, encontram-se transcritas 200.000 palavras em 18 localidades, 80.000 das quais, pertencentes a 7 localidades, já se encontram disponibilizadas na homepage do CLUL);
- etiquetagem morfológica da transcrição normalizada com recurso ao etiquetador automático desenvolvido pela equipa do projecto Tycho Brahe, da Universidade de Campinas (actualmente estão etiquetadas 140.000 palavras em 12 localidades, 80.000 das quais, pertencentes a 7 localidades, estão em vias de disponibilização);
- etiquetagem sintáctica da transcrição normalizada que segue o modelo definido pela equipa do projecto Penn-Helsinki Parsed Corpus of Middle English, da Universidade de Pennsylvania, com a necessária adaptação aos dados dialectais do português (esta etiquetagem está em curso e os dados serão disponibilzados no final do projecto).

A extensão final do *corpus* deste projecto está planeada para atingir as 500.000 palavras, estando previsto que no final de 2004 sejam atingidas as 200.000 palavras anotadas sintacticamente.

No âmbito deste projecto, encontram-se em elaboração duas teses de doutoramento e uma de mestrado. Uma obra sobre sintaxe dialectal do Português Europeu estará pronta para publicação no final de 2004.

6.2 Estudos das variantes flexionais do verbo, em Português continental falado (VarV)

Os principais objectivos deste projecto são: (i) fazer o levantamento das variantes da flexão verbal detectadas nas diferentes regiões dialectais do continente; (ii) estabelecer os padrões de flexão variantes observados, caracterizando-os sobretudo em termos de traços morfológicos e fonológicos, mas também do ponto de vista da relação entre os traços morfofonológicos e o comportamento sintáctico do agrupa-

mento 'verbo-clítico'; (iii) definir as principais áreas dialectais e geográficas de cada padrão flexional e o *continuum* flexional entre áreas; (iv) comparar os padrões flexionais do Português nortenho com os que se observam na zona fronteiriça da Galiza.

Até agora, no âmbito do projecto, onze localidades da rede portuguesa foram estudadas, tendo dado origem a três artigos, duas conferências e uma tese de mestrado.

Até 2004, a previsão é a seguinte: audição e selecção de excertos orais, com a respectiva transcrição ortográfica e fonética, em trinta pontos de inquérito; introdução dos materiais numa base de dados e análise linguística dos dados.

Passo a referir o tratamento que tem sido dado a alguns dos *corpora*.

Informatização dos materiais transcritos

Em 1994 foi desenhada uma base de dados de modo a permitir a informatização dos materiais constantes dos cadernos de inquérito do ALEPG.

O modelo relacional existente entre cada uma das tabelas que constituem essa base de dados torna possível a sua consulta sob múltiplos aspectos: relatórios das respostas obtidas em cada ponto da rede; relatórios das respostas obtidas para cada conceito na totalidade dos pontos da rede; relatórios dos conceitos sem resposta por inquérito; relatórios dos conceitos relacionados (conceitos que não fazem parte do questionário mas que estão relacionados com outros que integram o questionário). Além destes relatórios é também possível obter outros, tais como relatórios dos informantes; relatórios dos inquéritos já intoduzidos na base de dados e relatórios dos conceitos que constituem o questionário.

A base de dados permite outro tipo de consultas, nomeadamente o de delimitar geograficamente áreas lexicais ou de fenómenos fonéticos.

Um programa de cartografagem automática permite a elaboração de mapas a partir dos dados já introduzidos. Actualmente já se encontram introduzidos os materiais recolhidos nos Açores e na Madeira e cerca de 30 inquéritos do Continente.

Digitalização dos *corpora* orais

A totalidade das gravações do grupo da Variação encontra-se em suporte magnético analógico (fitas e cassettes) que, com o correr do tempo, se vai gradualmente deteriorando.

No programa plurianual da FCT de apoio às Unidades de Investigação foi aprovado, para o triénio 2000-2002, no âmbito dos fundos programáticos, o início de uma acção que visa a realização de uma cópia de salvaguarda em suporte digital de todos os *corpora* sonoros do CLUL que se encontrem em suporte magnético.

Relativamente aos *corpora* do grupo da Variação, até agora já foram digitalizadas cerca de 1750 horas de gravação em 1059 CD's. Paralelamente está a ser construída uma base de dados com o conteúdo de cada CD.

A partir da matriz de salvaguarda serão feitas cópias de trabalho que permitirão, no futuro, a constituição de *sub-corpora* devidamente tratados, quer na sua qualidade sonora, quer na sua temática.