O Arquivo Dialetal do CLUP: disponibilização *on-line* de um corpus dialetal do português

João Veloso[1,2] *& Pedro Tiago Martins*[2,3]
[1]Faculdade de Letras da Universidade do Porto
[2]Centro de Linguística da Universidade do Porto(*)
[3]Universidade de Barcelona

Abstract

The present work gives a general presentation of the Dialect Archive of the Center of Linguistics of the University of Porto, a database of spoken European Portuguese, which encompasses both mainland Portugal and the archipelagos of Madeira and Azores, effectively showcasing the dialectal variation of the language. The archive consists of a set of recordings collected during the last two decades, complete with full orthographic and narrow phonetic transcriptions, detailed maps and descriptions of the attested dialectal phenomena. In order to provide a general overview of the project, we go through its history, general aims, methodological concerns and materials.

Keywords: Dialect variation in Portuguese, dialectology, dialectal corpus, phonetic transcription
Palavras-chave: Variação dialetal do português, dialetologia, corpus dialetal, transcrição fonética

1. Breve apresentação do Arquivo Dialetal do CLUP

O Arquivo Dialetal do Centro de Linguística da Universidade do Porto é uma coleção de cerca de 140 amostras de fala do português europeu (PE) recolhidas desde 1994 por estudantes de linguística portuguesa da Faculdade de Letras da Universidade do Porto (FLUP) no âmbito de trabalhos curriculares de diversas disciplinas. A coleção encontra-se depositada nas instalações do Centro de Linguística da Universidade do Porto (CLUP), disponível para consulta por investigadores interessados no seu conteúdo, e parte dela (69 amostras) pode ser livremente consultada numa página *web* dedicada a este projeto (http://cl.up.pt/arquivo). Todas as amostras disponibilizadas na página do Arquivo e a maior parte das que se encontram em acesso mais reservado no CLUP foram sujeitas a transcrição ortográfica e fonética estreita e ao levantamento de fenómenos de variação dialetal registados numa base de

(*) Projeto Estratégico FCT: PEst-OE/LIN/UI0022/2011.

dados linguísticos. Todo o material que compõe este acervo se encontra devidamente catalogado. As amostras transpostas para a página *web* do Arquivo foram todas sujeitas a tratamento cartográfico. O trabalho de recolha de amostras e de organização do acervo continua presentemente em curso e em fase de ampliação.

Neste texto, daremos a conhecer os aspetos mais importantes da história e dos resultados deste projeto, bem como dos conteúdos da coleção e do tratamento linguístico e cartográfico a que ela foi submetida; apresentaremos também alguns dos desenvolvimentos futuros que se esperam.

2. História do Arquivo Dialetal do CLUP

Como foi afirmado na apresentação, o Arquivo Dialetal do CLUP tem as suas origens num conjunto de trabalhos curriculares solicitados, desde 1994, a estudantes de diversas disciplinas da área da linguística da Faculdade de Letras da Universidade do Porto no âmbito da avaliação curricular a que deveriam ser sujeitos sob a supervisão do primeiro autor deste artigo. No âmbito de tais trabalhos, era solicitado aos estudantes que recolhessem uma amostra gravada de uma produção do português europeu proveniente de um dialeto/socioleto em princípio diferente do do estudante que efetuava a recolha e que tal amostra fosse posteriormente objeto de um trabalho escrito que incluía a transcrição fonética e ortográfica do excerto sonoro, um pequeno comentário dialetal e outros elementos que enriquecessem a recolha (cartografia, documentação consultada, etc.). Os principais objetivos destes trabalhos escolares eram, em primeiro lugar, familiarizar os estudantes com algumas técnicas básicas de trabalho de campo em linguística e, simultaneamente, torná-los mais conscientes da questão da variação linguística.

Nos horizontes iniciais dessas propostas de trabalho não estava, portanto, a construção de um acervo de amostras dialetais que viesse um dia a ser disponibilizado ao público em geral. Este último objetivo surgiu posteriormente, quando a dimensão da coleção de amostras assim recolhidas atingiu um volume considerável, o que nos conduziu à sugestão do interesse que alguns dos seus dados poderia ter para um público mais vasto.

A ideia de, partindo do conjunto de recolhas efetuadas por estudantes da FLUP, construir um *corpus* estruturado de fala dialetal do português e de disponibilizá-lo publicamente ganhou um impulso importante quando, em 2008, foi possível passar a contar com a colaboração de

três estudantes a tempo parcial (em dois anos letivos sucessivos) para se dar início à organização, catalogação e verificação dos dados da coleção, graças ao programa de Bolsas BII (Bolsas de Iniciação à Investigação) da Fundação para a Ciência e a Tecnologia (FCT), ao abrigo de um pequeno projeto interno do CLUP a que foi dado o nome de *Fonotáticas.* Data dessa altura o início da colaboração do segundo autor deste artigo em todas as tarefas que dizem respeito à organização do Arquivo, colaboração mantida até ao presente sob diversas modalidades institucionais e contratuais.

Nessa primeira etapa de organização geral do Arquivo, a principal tarefa concretizada foi a organização de toda a coleção que viria a ser o seu fundo primitivo, até então dispersa, desorganizada e não catalogada. Neste sentido, foram inteiramente concluídas as seguintes tarefas, nessa fase inicial do projeto:

1) Inventariação e catalogação de todo o material recolhido;
2) Armazenamento desse material em condições próprias e seguras;
3) Verificação da qualidade de todas as gravações sonoras;
4) Verificação dos documentos escritos que acompanham as gravações.

Além da organização física de todo o material que integra a coleção original de dados do Arquivo, foi desta forma construída uma base de dados completa em que foram registadas em campos separados informações relativas à origem de cada amostra (ano e autor de cada recolha e proveniência geográfica e outros dados demográficos do informante).

Após o termo das bolsas BII da FCT, o projeto de construção do Arquivo foi interrompido entre 2010 e 2011. Em 2011, já com a designação oficial de *Arquivo Dialetal do Centro de Linguística da Universidade do Porto* e enquanto subprojeto definido dentro do programa geral de trabalho científico desta unidade de investigação, foi retomado com maior sistematicidade, novamente com a colaboração a tempo parcial do segundo autor do artigo e a coordenação do primeiro.

A reativação do projeto nesse ano trouxe consigo uma revisitação dos seus objetivos, sendo o ponto principal a preparação e divulgação de todo o acervo, dando-se especial importância ao desenvolvimento de recursos adicionais e à criação de uma página *web*, para fácil consulta por todos os interessados, de forma completamente livre. Foi também tida em

conta a chegada de novas amostras, obtidas na mesma modalidade das amostras recolhidas até então. O novo plano de trabalho foi dividido em duas partes: a primeira dedicada à preparação do material e a segunda à sua divulgação. Da primeira parte fizeram parte os seguintes pontos:

1) Nova verificação de todo o material catalogado;
2) Catalogação do material novo;
3) Transcrição ortográfica de todo o material;
4) Transcrição fonética estreita de todo o material;
5) Alinhamento de transcrições ortográficas e fonéticas;
6) Levantamento exaustivo de marcas dialetais e fenómenos fonéticos e/ou fonológicos

Após a conclusão das tarefas elencadas acima, foi dado início a uma nova fase do projeto, a que correspondem os seguintes pontos:

1) Preparação de todo o material para divulgação;
2) Tratamento cartográfico exaustivo de todo o material;
3) Produção de documentação técnica;
4) Desenvolvimento e criação de uma página *web*, com os seguintes objetivos:
5) Disponibilização de gravações sonoras;
6) Disponibilização de informação sociodemográfica dos informantes;
7) Disponibilização de transcrições ortográficas e fonéticas;
8) Disponibilização de informação cartográfica;
9) Disponibilização de informação linguística;
10) Apresentação de todo este material de forma facilmente acessível.

Nas secções 5 e seguintes, faremos uma descrição sucinta destas duas fases, tentando realçar as preocupações metodológicas associadas a cada aspeto específico do projeto.

3. Metodologia das recolhas e características gerais das amostras

Como foi referido, todas as amostras do Arquivo têm a sua origem em trabalhos escolares realizados por estudantes de linguística da FLUP. Tais trabalhos, de natureza facultativa ou obrigatória consoante as diversas disciplinas e anos letivos, deveriam obedecer às seguintes instruções e características:

- cada estudante deveria recolher uma entrevista espontânea, de tema livre, com a duração aproximada de 20 a 30 minutos, com um falante nativo do português europeu proveniente de uma norma dialetal diferente da do entrevistador;

- a entrevista deveria ser gravada num suporte à escolha do entrevistador, de entre os meios ao seu alcance;

- do total da entrevista, o estudante deveria escolher uma pequena porção com uma duração de entre 60 a 90 segundos, que parecesse ao entrevistador particularmente rica em termos de variação dialetal; o trabalho final deveria incidir exclusivamente sobre este segmento da entrevista;

- os elementos a entregar ao professor seriam a gravação propriamente dita da porção de 60-90 s de fala e um pequeno trabalho escrito com o seguinte conteúdo: identificação sociodemográfica do falante (idade, sexo, escolaridade, profissão, origem geográfica e local da recolha), transcrição ortográfica e fonética do excerto gravado, identificação explícita dos fenómenos de variação atestados na gravação, breve comentário dialetal, com identificação das principais marcas que permitissem a inclusão do falante numa área dialetal definida, bibliografia consultada, anexos e outros elementos (mapas, ilustrações, etc.).

Esta metodologia trouxe vantagens e desvantagens: por um lado, a pouca exigência técnica das recolhas – os estudantes podiam escolher livremente a origem dialetal dos seus informantes e deveriam usar os meios técnicos de gravação ao seu alcance, em função das suas disponibilidades e conveniências pessoais – fez com que muitos estudantes tivessem aderido à proposta de trabalho e, assim, o número de amostras recolhidas tenha sido bastante significativo; por outro lado, porém, a qualidade sonora das amostras gravadas e a fiabilidade de muitas das informações contidas nos trabalhos nem sempre se revelaram as mais adequadas a um trabalho de investigação científico na área da linguística. Relativamente a este último ponto, recordamos mais uma vez que nunca esteve nos objetivos iniciais das propostas feitas aos estudantes a construção de um recurso científico como o Arquivo e que boa parte do trabalho feito no âmbito deste projeto consistiu precisamente em tentar colmatar, na medida do possível, algumas falhas e lacunas dos trabalhos constantes da coleção, nomeadamente através de uma seleção criteriosa das amostras, do tratamento acústico de algumas delas, da verificação sistemática de toda a informação e da transcrição de raiz de todo o material (ignorando completamente as transcrições feitas pelos estudantes autores das recolhas).

Algumas das recolhas feitas pelos estudantes, revelando-se completamente inaproveitáveis para um tratamento linguístico minimamente fiável, acabaram mesmo por ser excluídas do acervo geral do Arquivo.

4. Caracterização geral do material do Arquivo

Com base na situação descrita na secção anterior foi possível recolher um conjunto de 141amostras. 56 amostras são produzidas por informantes do sexo masculino e 85 por falantes do sexo feminino. A origem geográfica dos falantes é a que se encontra no Quadro 1.

Aveiro	14	Portalegre	1
Braga	11	Porto	49
Bragança	5	Santarém	1
Coimbra	3	Setúbal	1
Faro	4	Viana do Castelo	6
Guarda	1	Vila Real	9
Leiria	4	Viseu	6
Lisboa	7	Açores	5
		Madeira	7
		Desconhecido	7
		Distritos sem amostras recolhidas:	Beja, Castelo Branco, Évora

Quadro 1: Origem geográfica das amostras (por distrito/região autónoma), no total das amostras do Arquivo (N=141)

OBS.: Distrito/região autónoma onde os inquiridos, segundo informação constante do trabalho escrito entregue pelos autores das recolhas, haviam passado a maior parte da sua vida

Conclui-se assim que a maior parte dos dialetos representados no acervo do Arquivo correspondem aos dialetos setentrionais, com particular destaque para a zona subdialetal do Baixo Minho e Douro Litoral.

Destas 141 amostras, só cerca de metade foi selecionada para figurar na página *web* do Arquivo. As amostras excluídas foram-no, na maior parte dos casos, devido à qualidade acústica muito deficiente, por vezes no limite da inteligibilidade, das gravações sonoras. Outros motivos que levaram à exclusão de algumas amostras foram a extrema facilidade de identificação civil dos informantes ou o facto de a recolha ter sido baseada na modalidade de leitura a partir de texto escrito.

As 69 amostras que correspondem à parte do Arquivo que está disponível na página *web* provêm de 34 informantes do sexo masculino e 35 do sexo feminino e repartem-se geograficamente de acordo com os dados encontrados no Quadro 2.

Aveiro	6	Portalegre	1
Braga	10	Porto	18
Bragança	3	Santarém	0
Coimbra	2	Setúbal	0
Faro	3	Viana do Castelo	4
Guarda	0	Vila Real	3
Leiria	2	Viseu	6
Lisboa	5	Açores	2
		Madeira	4
		Distritos sem amostras recolhidas:	Beja, Castelo Branco, Évora

Quadro 2: Origem geográfica das amostras (por distrito/região autónoma), no subtotal das amostras do Arquivo disponibilizadas nesta página (N=69) 6+10

OBS.: Distrito/região autónoma onde os inquiridos, segundo informação constante do trabalho escrito entregue pelos autores das recolhas, haviam passado a maior parte da sua vida

5. Tratamento do material

Cada amostra pertencente ao Arquivo foi sujeita a um tratamento completo cujo principal objetivo foi garantir o seu aproveitamento, com um grau satisfatório de qualidade e fiabilidade, para o estudo linguístico ou para a ilustração minimamente validada de fenómenos de variação específicos.

5.1 Verificação

De modo a garantir a fiabilidade do material constante do acervo do Arquivo, cuja catalogação, tanto física como informatizada, foi iniciada ainda na fase do projeto *Fonotáticas,* fez-se uma verificação inicial de cada amostra. Em termos práticos, esta verificação preliminar visou garantir que o material catalogado fisicamente correspondia ao catalogado digitalmente, assim como corrigir quaisquer imprecisões no tocante à caracterização de cada amostra.

5.2 Catalogação

A catalogação de cada amostra consistiu em atribuir um número de inventário a cada amostra e ao seu registo numa base de dados de que constam a informação sociodemográfica do falante, as coordenadas geográficas da recolha e as principais marcas de variação dialetal atestadas na amostra. Trata-se de um trabalho não inteiramente circunscrito no tempo, visto que a coleção do Arquivo se encontra em permanente expansão, tendo em atenção que

continua a ser proposto aos estudantes trabalho desta índole. O número de amostras presentemente catalogadas ascende a 141.

5.3 Transcrição

A fase nuclear do tratamento de cada amostra residiu na sua transcrição (ortográfica e fonética); este é o passo mais importante em todo o projeto, que transforma as gravações sonoras em material utilizável e sobre o qual se podem facilmente efetuar pesquisas específicas. Sem estarem transcritas, as amostras não passam de ficheiros de som, interessantes a título de curiosidade mas pouco úteis a um linguista que necessite de dados linguísticos ou de material preparado para estudar. O trabalho de transcrição realizado é descrito nas secções abaixo.

5.3.1 Transcrição ortográfica

Num primeiro momento, cada amostra, depois da verificação prévia e da catalogação acima descritas, foi sujeita a transcrição ortográfica.

O processo utilizado para a transcrição ortográfica foi relativamente simples: os ficheiros de som são reproduzidos num computador e aquilo que é ouvido é escrito num ficheiro de texto. Não são usadas pontuação, maiúsculas, abreviaturas, numerais ou outras convenções ortográficas, sendo as gravações transcritas palavra por palavra, separando-se cada uma por espaços em branco de acordo com as divisões ortográficas canónicas da língua.

A transcrição ortográfica permitirá no futuro efetuar facilmente pesquisas por palavra ou por expressão. É ainda possível que, no futuro, possa ser feita sobre o material etiquetagem morfossintática, ou qualquer outro tratamento não fonético; tais tarefas beneficiarão do facto de o material estar já transcrito na íntegra.

A maior dificuldade relacionada com a transcrição ortográfica é a qualidade das gravações, algumas das quais no limite da inteligibilidade. Em alguns casos, foi necessário recorrer à reprodução dos ficheiros de som a velocidades menores, para transcrever algumas das gravações que despertaram dúvidas em relação ao seu conteúdo.

5.3.2 Transcrição fonética

Todas as amostras foram sujeitas a transcrição fonética estreita, na qual tivemos o objetivo de registar todos os detalhes de realização fonética auditivamente percetíveis, independentemente de terem ou não relevância dialetal. Mais adiante, daremos algumas

indicações sobre a forma como se desenrolou o processo de transcrição fonética de todas as amostras. A transcrição fonética é, com efeito, a mais importante transcrição efetuada sobre o material. Trata-se de um processo moroso e muito dependente da perceção de quem transcreve. Disponibilizar uma base de dados com amostras dialetais transcritas foneticamente permite a qualquer interessado fazer buscas por fenómenos e por processos específicos e dessa procura retirar dados úteis. Além disso, será possível a longo prazo estabelecer um mapa dialetal atualizado, tão fiável quanto a qualidade das gravações e das respetivas transcrições.

O processo utilizado para transcrição fonética foi, naturalmente, mais complexo do que aquele usado para a transcrição ortográfica. Numa primeira fase optou-se por uma transcrição manual, feita a partir da audição direta dos ficheiros de som. Depois, as transcrições manuais foram passadas para ficheiros de texto, podendo então ser mais facilmente manipuladas. A duração média do processo de obtenção de uma primeira transcrição fonética completa é de cerca de 2 horas por cada minuto de gravação. Tal como acontece com a transcrição ortográfica, a natureza de cada gravação influencia bastante esta duração.

Todas as transcrições foram realizadas pelo segundo autor e repetidamente revistas e melhoradas por ambos, com base numa análise auditiva e, nos casos considerados dúbios ou por natureza difíceis de transcrever, com o auxílio de *software* de análise acústica (Praat[1]).

Um número significativo de transcrições foi ainda sujeito a uma experiência de *Inter-Judge Agreement* (Shriberg *et al.*, 1984; Shriberg & Lof, 1991) para se fazer a avaliação da adequação da transcrição de aspetos problemáticos das realizações fonéticas, tendo sido escolhidos quatro fenómenos em que o número de transcrições dúbias, não resolvidas pela revisão feita pelos autores nem pela análise acústica do material, atingiu um volume considerado representativo (Martins & Veloso, 2012): realização/não realização do chevá, diferentes pontos e modos de articulação da "vibrante múltipla", fricatização/não fricatização das oclusivas sonoras intervocálicas e realização vozeada/não vozeada de sílaba final átona. Um painel de 7 transcritores (investigadores e estudantes pós-graduados de fonética e linguística e terapeutas da fala, todos com treino fonético) ouviu 20 amostras selecionadas e fez a respetiva transcrição fonética estreita. Foram consideradas como definitivas as transcrições com uma percentagem de concordância entre transcritores igual ou superior a 75%.

[1] http://www.fon.hum.uva.nl/praat/

A metodologia seguida para a realização e verificação das transcrições fonéticas de todo o material do Arquivo confere-lhes assim, em nossa opinião, um elevado grau de segurança e fiabilidade; acompanhando os ficheiros sonoros e restantes conteúdos, estas transcrições fonéticas constituem, deste modo, uma fonte de informação muito importante para todos os leitores da página *web* do Arquivo

Tentámos atingir um nível de detalhe compatível com a transcrição fonética alofónica e comparativa, de acordo com a tipologia de Ladefoged (1993). À exceção do acento de palavra em palavras lexicais e palavras gramaticais com mais de duas sílabas, não foi considerado nenhum outro processo fonológico lexical sem reflexo direto ou sistemático no sinal acústico. Por facilidade de leitura, optou-se por separar por espaços em branco os segmentos de transcrição correspondentes às palavras ortográficas. Algumas convenções e símbolos fonéticos utilizados poderão ser pouco familiares relativamente a descrições anteriores do português. De entre estes, destacam-se os seguintes:

- símbolos para os **róticos**: foram identificadas sete realizações diferentes das consoantes normalmente designadas na literatura por "vibrantes", correspondentes aos segmentos designados também, noutras propostas terminológicas, por róticos: vibrante múltipla alveolar [r], vibrante múltipla uvular [ʀ], fricativa uvular vozeada [ʁ], fricativa velar desvozeada [x], fricativa uvular desvozeada [χ], vibrante simples alveolar [ɾ] e vibrante simples retroflexa [ɽ];

- **lateral retroflexa**: foi identificada em algumas amostras uma realização retroflexa da lateral alveolar, tendo sido utlizado nesses casos o símbolo correspondente à aproximante lateral retroflexa ([ɭ]);

- **desvozeamento**: foram assinaladas com o diacrítico de desvozeamento (◌̥ ou ◌̬) todas as vogais e consoantes fonologicamente vozeadas mas foneticamente realizadas sem vozeamento, mesmo quando encontradas em contextos não referidos por estudos anteriores;

- **fricativas alveolares**: a distinção pré-dorsal/apical nas fricativas alveolares foi sistematicamente assinalada através do uso dos símbolos correspondentes à fricativa retroflexa vozeada ([ʂ]) e fricativa retroflexa vozeada ([ʐ]) nos casos de realização apical;

- **coarticulação de vogais**: foram identificados alguns casos a que aqui nos referiremos como coarticulações vocálicas. O fenómeno, atestado especialmente nos dialetos setentrionais, consiste na mudança da qualidade vocálica de uma vogal (normalmente [o], mas não

exclusivamente esta) na sua fase final, sendo a realização final mais próxima de [ɐ]. Este fenómeno foi assinado através de uma ligadura unindo as duas vogais coarticuladas (p. ex.: [o͡ɐ]; [e͡ɐ]);

- **ausência de consoantes glotais**: há várias ocorrências da consoante glotal não vozeada ([ʔ]), cujo registo seria de esperar dado o nível de rigor das transcrições. No entanto, por ser extremamente difícil de detetar inequivocamente – e também por não ser um alofone sistematicamente atestado –, optou-se por omitir este segmento na grande maioria dos casos.

Na página do Arquivo pode ser consultada uma lista integral de todos os símbolos fonéticos e diacríticos utlizados.

5.3.3 Transcrição alinhada

A partir das transcrições ortográfica e fonética de cada amostra, foi criado para cada ficheiro um documento novo em que se fez um alinhamento em pares de linhas, palavra ortográfica a palavra ortográfica, das duas transcrições (v. exemplo em anexo). Com este tipo de transcrição, os leitores da página *web* do Arquivo podem dispor de um instrumento de consulta que tornará mais clara a natureza lexical e fonética do acervo e possibilitará a busca de realizações fonéticas a partir das entradas lexicais de toda a coleção.

5.4 Levantamento de marcas dialetais

As transcrições descritas nas secções anteriores, com a verificação cruzada a que foram sujeitas, permitiram fazer um levantamento muito objetivo dos fenómenos de variação atestados em cada amostra.

Os fenómenos considerados neste levantamento foram os seguintes:

- desvozeamento de vogais átonas (em posição pré-tónica e pós-tónica);
- realização de vogais e ditongos nasais (em posição tónica);
- realização de ditongos orais;
- fricatização de consoantes oclusivas vozeadas em posição intervocálica;
- realização de /v/ como [ʋ];
- realização de róticos ([r], [ʀ], [ʁ], [x], [χ], [ɾ] e [ɽ]);
- realização da lateral /l/;
- realização de /S/ diferente de [z] em final de palavra antes de vogal;
- realização de sibilantes apicais;

- aspiração de consoantes oclusivas;
- manutenção ou neutralização da oposição /v/~/b/;
- realização de africada palatal não vozeada;
- ausência de redução do vocalismo átono.

Este levantamento contemplou ainda a realização de vários segmentos problemáticos ou que na nossa opinião se revestem de interesse para a análise da variação em português europeu. No total, foi registada numa base de dados interna informação de variação dialetal repartida por 63 variáveis diferentes. Destas, uma grande parte foi transposta para as descrições linguísticas das amostras disponibilizadas na página do Arquivo. Tal como no caso das transcrições fonéticas, o levantamento de marcas dialetais passou por várias fases de revisão e validação, a fim de se garantir o maior rigor e fiabilidade possíveis.

Para a escolha destes fenómenos concorreram essencialmente três fatores principais:

- primeiro, alguns deles são os contemplados pelos anteriores levantamentos dialetais gerais do português europeu, nomeadamente por Cintra (1971) para a determinação das fronteiras entre dialetos setentrionais e dialetos centro-meridionais da língua;

- em segundo lugar, dado que muitas das amostras do Arquivo são provenientes da zona do Grande Porto e concelhos limítrofes, foi também prestada especial atenção, na análise e organização dos dados, às marcas dialetais que Cintra (1971) considera típicas da zona subdialetal do Douro Litoral e Baixo Minho, tais como as realizações de <ão> e o abaixamento de vogal central tónica em contexto nasal;

- finalmente, e num plano não menos importante, a própria observação empírica dos dados tratados, aprofundada pela transcrição fonética estreita de todas as amostras, mostrou que estes correspondiam aos casos mais evidentes de variação. Sublinhe-se aqui que alguns destes fenómenos, tanto quanto nos é dado conhecer, não se encontram suficientemente descritos por investigação anterior (p. ex., a realização de /v/ como [ʋ] ou a aspiração de oclusivas surdas), o que contribui, em nosso entender, para o interesse e originalidade deste acervo[2].

Na página do Arquivo pode ser consultada uma lista exaustiva de todas as marcas dialetais e fenómenos fonéticos e/ou fonológicos recolhidos, sendo feito também um contraste

[2] Nas transcrições fonéticas, encontram-se registados muitos outros casos de variação que não cabem nestas categorias e que não foram contemplados neste levantamento final por corresponderem a fenómenos menos frequentes ou menos sistemáticos.

com descrições anteriores da língua. Uma componente verdadeiramente essencial deste levantamento encontra-se no tratamento cartográfico a que ele foi sujeito. Este tratamento esteve a cargo de Miguel Nogueira, da Oficina do Mapa da FLUP, que orientou uma parte importante da organização geral de toda a informação linguística recolhida, num verdadeiro trabalho interdisciplinar e em equipa entre linguistas e geógrafos com vista ao correto tratamento cartográfico da informação linguística de que dispúnhamos. Todos os mapas assim gerados, bem como os critérios para a conceção dos mesmos, se encontram na página *web* do Arquivo (v. exemplo em Anexo) e oferecem uma ilustração muito clara das variações atestadas.

O cruzamento de todas as variáveis linguísticas e extralinguísticas registadas no trabalho em torno do Arquivo dará origem a breve trecho, segundo as nossas presentes expectativas, a uma ferramenta de pesquisa dentro da própria página que permitirá aos leitores a localização rápida e segura de amostras concretas em que sejam atestados determinados fenómenos ou observadas certas características dos falantes. O trabalho prévio que conduzirá a esta facilidade na página, em nosso entender, encontra-se atualmente concluído, graças a todo o trabalho de catalogação, transcrição e levantamento de marcas linguísticas feito ao longo dos últimos anos.

5.5 Edição de gravações

Uma das maiores dificuldades com que o projeto se deparou foi a variedade de formatos em que se encontram as amostras. Esta variedade não é de estranhar, na medida em que as amostras foram recolhidas ao longo de quase duas décadas, começando no ano letivo de 1994/1995, altura em que dominavam os formatos analógicos de armazenamento de áudio, como a cassete e a minicassete. A fim de garantir a qualidade das gravações, mas também para facilitar o seu manuseamento, as amostras neste formato estão a ser submetidas a um processo de digitalização. No trabalho de transcrição e levantamento linguístico, foi dada prioridade às amostras já em formato digital, tentando sempre manter um grupo de amostras representativo da variação dialetal do português europeu. Todas as gravações foram editadas tendo em consideração dois aspetos: (i) *o conteúdo*, tendo sido feitos esforços para eliminar elementos reveladores da identidade dos informantes ou terceiros, elementos de outra forma comprometedores, elementos violadores de direitos de autor, intervenções excessivas de terceiros, ruídos e pausas longas, e (ii) a *qualidade técnica* das gravações, tendo sido levado a

cabo um trabalho de equalização, nivelação e compressão das amostras, individualmente e entre si.

5.6 Tratamento cartográfico

Como já foi referido, em colaboração com a Oficina do Mapa da Faculdade de Letras da Universidade do Porto, todo o conteúdo das amostras do Arquivo foi devidamente cartografado. Deste trabalho, resultaram mapas de grande qualidade técnica de acordo com os padrões atuais da cartografia científica.

Estes mapas estão acessíveis na página *web* do Arquivo e dividem-se essencialmente por três tipos:

- mapas de caracterização geral da população, com a repartição espacial dos inquiridos repartidos por características sociodemográficas (sexo, escolaridade, etc.);

- mapas de variação linguística, com a distribuição geográfica, neste acervo, dos fenómenos de variação linguística atestados (cada um dos mapas deste tipo faz-se acompanhar de uma gravação isolada do fenómeno relevante);

- mapa de cada amostra, com a localização exata do local em que a gravação foi recolhida.

Na página *web* do Arquivo, são dadas informações relevantes sobre os critérios de feitura dos mapas, os quais podem condicionar, em vários aspetos, a interpretação geográfica dos mesmos. Ressalve-se que as zonas do território cobertas pelas recolhas não se encontram unanimemente distribuídas: existe uma clara prevalência das recolhas setentrionais e próximas de centros urbanos.

6. Disponibilização do material: a página *web* do Arquivo

A página *web* do Arquivo (www.cl.up.pt/arquivo) foi criada com a finalidade de disponibilizar todo o material e facilitar a sua consulta. Por enquanto, o projeto conta com uma página estática, ou seja, todos os materiais estão organizados numa espécie de repositório, podendo ser consultados individualmente. No entanto, está planeado o desenvolvimento de uma ferramenta de pesquisa, através da qual será possível fazer pesquisas por parâmetros específicos, fazendo cruzamento de dados sociodemográficos e linguística. Será possível procurar, por exemplo, por falantes de uma faixa etária específica que realizem um ditongo de uma determinada maneira e não de outra.

Nas secções abaixo é descrito o modo como os materiais são disponibilizados.

6.1 Gravações

Cada amostra pode ser escutada diretamente na página do Arquivo, bastando para isso utilizar um navegador compatível. Neste sentido, de modo a responder às particularidades dos navegadores mais utilizados, foi feita uma tentativa de providenciar as gravações em vários formatos (*mp3, ogg* e *wav*), que cada navegador selecionará de modo automático, conforme as suas especificações.

6.2 Informação sociodemográfica

Numa secção própria, é possível consultar vários quadros relativos a informação sociodemográfica, estando de momento disponíveis dados sobre o sexo e proveniência dos falantes. Embora não disponibilizados na versão atual da página, foram também recolhidos dados acerca da escolaridade e idade dos informantes.

6.3 Descrição linguística

Cada amostra faz-se acompanhar de uma breve descrição linguística da variedade do falante, com base no levantamento de marcas dialetais feito anteriormente, e utilizando as mesmas convenções aplicadas nas transcrições fonéticas. Optou-se por incluir na descrição apenas marcas dialetais positivas, ou seja, a "não realização" de um segmento ou ditongo não é normalmente contemplada. Da ausência de uma marca dialetal em cada descrição pode inferir-se que não foi atestada.

7. Resultados

Além da organização e catalogação completa de todo o Arquivo propriamente dito, da transcrição de todo o material e da construção da página, com todos os conteúdos dela constantes, o trabalho em torno deste projeto originou, direta ou indiretamente, outros resultados importantes, de que aqui salientamos:

- a utilização de dados do Arquivo em pesquisas conduzidas, de momento ainda inéditas, por investigadores do Brasil e da Finlândia;

- a realização de algum trabalho empírico, como a já referida exploração experimental da fiabilidade das transcrições fonéticas do material com base na metodologia do *Inter-Judge Agreement*;

- a apresentação de resultados em algumas jornadas e encontros científicos (II e III Jornadas de Estudos Pós-Graduados em Fonologia da FLUP, 45º Congresso Anual da Societas Linguistica Europaea, XXVIII Encontro Nacional da Associação Portuguesa de Linguística – no caso destes dois últimos eventos, dando origem à apresentação de um póster [3] e à publicação desta comunicação);

- a melhoria do protocolo de recolha de novas amostras, com regras mais claras e uniformes para a gravação das amostras sonoras e a identificação sociodemográfica e geográfica dos falantes e com obtenção de consentimento explícito para a recolha e para o seu aproveitamento para o acervo do Arquivo.

8. Desenvolvimento futuro

Nesta secção final do artigo, daremos conta dos passos de desenvolvimento do projeto do Arquivo Dialetal do CLUP atualmente em curso ou com concretização prevista para as fases seguintes.

Expansão da coleção

O Arquivo Dialetal do CLUP é um projeto em crescimento. Será dada continuidade aos trabalhos de recolha que, nas últimas duas décadas, permitiram acumular as gravações que formam agora parte do seu acervo. No entanto, as diretrizes a seguir para a recolha de amostras estão neste momento mais bem delineadas (vd. secção anterior), tendo sido melhorados os critérios técnicos e de registo de informação sociodemográfica de acordo com as necessidades do projeto.

Revisão e acertos

Durante o desenvolvimento do projeto, foram feitos esforços no sentido de se garantir a fiabilidade do material disponibilizado. Esta preocupação mantém-se por natureza sempre presente, e consideramos que o Arquivo está sempre sujeito a novas revisões e acertos, tarefas que continuarão a ser levadas a cabo pelos colaboradores do projeto.

[3] Martins & Veloso (2012): http://cl.up.pt/arquivo/artigos/Poster_SLE.pdf

Etiquetagem morfossintática

Como já foi referido, o *corpus* recolhido ascende atualmente a cerca de 140 amostras de produção espontânea oral em português. Até ao presente momento, o estudo deste material cingiu-se a aspetos da sua realização fonética e variação dialetal. Estamos em crer, porém, que a riqueza deste material pode vir a revelar-se útil para outros domínios de estudo sobre as estruturas e a variação do português e para o conhecimento da sua evolução nas últimas duas décadas. Pensamos, por isso, que será possível alargar o âmbito da análise deste material linguístico a domínios como a organização sintática e discursiva e que, para tal, poderá vir a submeter-se todo o material a um trabalho de etiquetagem morfossintática.

Dicionário

Ainda no âmbito do alargamento do estudo deste *corpus* a outras áreas que não a das realizações fonéticas e variação dialetal, está nos horizontes do desenvolvimento do projeto a construção de um glossário com todas as entradas lexicais de palavras realizadas no *corpus,* os seus índices de frequência e a localização exata de cada qual, nas suas várias ocorrências, no *corpus* total.

Ferramenta de pesquisa

Como foi referido na secção 6, está previsto o desenvolvimento de um instrumento de pesquisa dentro da própria página *web* do Arquivo, que permitirá aos utilizadores efetuar procuras específicas sobre as amostras, criando-se a possibilidade de cruzar dados sociodemográficos e linguísticos. Esta nova funcionalidade será um passo em frente, abrindo-se o caminho para usos mais sofisticados e próximos das necessidades dos interessados na linguística e dialetologia do português.

Referências

Cintra, L. F. L. (1971) Nova proposta de classificação dos dialectos galego-portugueses. *Boletim de Filologia* 22, pp. 81-116.

Ladefoged, P. (1993) *A course in phonetics*. Fort Worth, TX: Harcourt, Brace and Jovanovich.

Martins, P. T. & Veloso J. (2012) *Inter-judge agreement in transcribing dialectal data: a study of a corpus of dialectal Portuguese.* Poster apresentado na 45th Annual Meeting of the Societas Linguistica Europaea, Universidade de Estocolmo. http://cl.up.pt/arquivo/artigos/Poster_SLE.pdf

Shriberg, L. D., Kwiatkowski, J. & Hoffmann, K (1984) A procedure for phonetic transcription by consensos. *Journal of Speech and Hearing Research* 27, pp. 456-465.

Shriberg, L. D. & Lof, G. L. (1991) Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics* 5(3), pp. 225-279.

Anexos

Anexo1: Exemplo de uma transcrição alinhada (Amostra 27)

vim para o porto vai fazer quatro anos a primeira experiência foi um bocado chata
vĩ p ɔ ˈpoɾtu̯ βaj feˈze ˈkwatɾw ˈanuʃ ɐ pɾiˈmejɾɐ ʃˈpɾj̃ẽsjɐ foj ũ bˈkaɖ ˈʃatɐ
porque porque tinha dezassete anos na altura e foi a primeira vez que saí de casa
ˈpuɾkɨ pu̯ɾ ˈtiɲɐ dzɐˈsɛt ˈanuʃ n aɬˈtuɾɐ ii foj ɐ pɾi̯ˈmejɾɐ veʃ k sɐˈi d kaz
assim e depois numa cidade nova sem os pais longe dos amigos foi um bocadinho
aˈsĩ i dˈpojʃ ˈnumɐ siˈdad ˈnɔβɐ sẽj̃ŋ uʃ pajʃ lõʒ duz ɐˈmiguʃ foj ũ bkɐˈðiɲ
complicado também fui morar para uma casa que era tipo uma residencial com
kũpliˈkaɖ tɐˈmẽj̃ fuj muˈɾa pa wmɐ ˈkazɐ kʲ ˈɛɾɐ tip umɐ ʁizidẽˈsjɐɬ kõ
muita gente e não foi assim muito agradável mas depois habituei-me agora estou
ˈmũj̃tɐ ˈʒẽtɨ ii nẽw̃ foj ɐˈsĩ ˈmũj̃t ɐgɾɐˈðaβɛɬ mɐʒ dˈpojz ɐβiˈtwɐjm ɐˈɣɔɾɐ ʃto
totalmente integrada as coisas que eu gostei adoro a cidade acho que agora já não
tu̯taɬˈmẽt ĩtˈgɾadɐ̯ ɐʃ ˈkojzɐʃ k ew ɣʃˈtɐj ɐˈðɔɾ ɐ siˈðaɖi̯ aʃ k ɐˈɣɔɾɐ ʒa nẽw̃
ia viver para a minha terra que é ponte de lima sei lá de monumentos eu já
ˈiɐ viˈve p a ˈmiɲɐ ˈtɛχɐ k ɛ põt d ˈlimɐ sej la d munuˈmẽtuʒ ew ʒa
conhecia o porto mas gosto gosto muito da baixa a foz também é muito bonita mas
kuɲɨˈsi ɔ ˈpoɾtu mɐʃ ˈgɔʃtu ɣɔʃt mũj̃ dɐ ˈbajʃɐ ɐ fɔʃ tɐˈmẽj̃ ɛ mũj̃t bˈnitɐ mɐʃ

Anexo 2: Exemplo de um dos mapas do Arquivo.
Distribuição geográfica da realização de <ou> como [ow]