Revista RBBA

Revista Binacional Brasil Argentina

ATLAS SINTÁTICO DO PORTUGUÊS EUROPEU: UM RECURSO EM CONSTRUÇÃO

ATLAS SINTÁCTICO DEL PORTUGUÉS EUROPEO: UN RECURSO EN CONSTRUCCIÓN

Catarina Magro
Universidade de Lisboa – ULISBOA
cmagro@letras.ulisboa.pt

Gael Vaamonde
Universidad de Granada – UGR
gaelvaamonde@ugr.es

Resumo
Este artigo apresenta um projeto em curso que visa conceber, construir e disponibilizar em linha um atlas digital da sintaxe dos dialetos do Português Europeu (SynAPse). Este atlas combina um corpus dialetal sintaticamente anotado, um motor de busca sintática e uma aplicação de webGIS e permite mapear automaticamente e de forma dinâmica os resultados de pesquisas sintáticas definidas pelos utilizadores, evidenciando a correlação espacial entre fenómenos e facilitando a identificação no território português de áreas geográficas de convergência sintática. Neste artigo, discutem-se as vantagens de um recurso cartográfico com estas características e funcionalidades para a investigação em sintaxe dialetal. Particularmente, defende-se a sua relevância para a exploração da dimensão espacial da variação sintática e para uma análise comparada das variedades dialetais numa perspetiva de Princípios & Parâmetros. Descreve-se ainda o plano de implementação técnica do atlas, com particular enfoque na etapa do processo atualmente concluída: a edição digital em XML-TEI do corpus de discurso dialetal que alimenta a ferramenta a construir e sustenta empiricamente a investigação a desenvolver.

Palavras-chave: Sintaxe dialetal. Dialetos do português europeu. Corpora digitais. Mineração de dados sintáticos. Cartografia linguística.

Resumen
Este artículo presenta um proyecto en curso cuyo objetivo es concebir, construir y ofrecer en línea un atlas digital sobre la sintaxis de los dialectos del portugués europeo (SynAPse). Este atlas combina un corpus dialectal sintácticamente anotado, um motor de búsqueda sintáctica y una aplicación webGIS, y permite mapear automáticamente y de forma dinámica los resultados de búsquedas sintácticas definidas por los usuarios, evidenciando la correlación espacial entre fenómenos y facilitando la identificación en el territorio portugués de áreas geográficas de convergencia sintáctica. En este artículo, se exponen las ventajas de un recurso cartográfico con estas características y funcionalidades para la investigación en sintaxis dialectal. Particularmente, se defiende su relevancia para la exploración de la dimensión espacial de la variación sintáctica y para un análisis comparado de las variedades dialectales desde la perspectiva de Principios y Parámetros. Finalmente, se describe el plano de implementación técnica del atlas, con especial atención a la etapa del proceso que está actualmente concluida: la edición digital en XML-TEI del corpus de discurso dialectal que alimenta a la herramienta aquí descrita y sustenta empíricamente la investigación que con ella se pretende desarrollar

Palabras clave: Sintaxis dialectal. Dialectos del portugués europeo. Corpus digitales. Minería de datos sintácticos. Cartografía lingüística.

1- Introdução

Este artigo dá a conhecer um novo projeto de investigação na área da sintaxe dialetal em desenvolvimento no Centro de Linguística da Universidade de Lisboa (CLUL). O projeto SynAPse – Syntactic Atlas of European Portuguese (Atlas Sintático do Português Europeu) pretende (i) investigar a dimensão espacial da variação sintática em português europeu (PE) a partir de uma perspetiva que combina a abordagem da sintaxe comparada e a metodologia geolinguística e (ii) criar um recurso cartográfico digital para visualizar a distribuição geográfica de traços sintáticos no espaço do território português e identificar áreas dialetais de base sintática[i].

Com uma duração prevista de três anos, este projeto teve início em outubro de 2018 e concluiu recentemente a primeira etapa do seu programa de trabalhos: a edição digital em XML-TEI do corpus dialetal que está na base do recurso cartográfico a construir e que suporta empiricamente a investigação a desenvolver. A descrição das decisões e estratégias relativas à

codificação em XML-TEI dos dados do CORDIAL SIN – Corpus Dialetal para o Estudo da Sintaxe (MARTINS, 2000) é, pois, a matéria principal do presente artigo e a secção 4 é-lhe dedicada. As secções que a antecedem apresentam a fundamentação, os objetivos e a metodologia do projeto. Assim, na secção 2, discute-se a importância de considerar a expressão areal da variação nos estudos em sintaxe dialetal; na secção 3, descrevem-se em detalhe os objetivos do projeto e a abordagem planeada para o desenvolvimento tecnológico do atlas e para a análise linguística e espacial dos dados. A secção 5 encerra o artigo com considerações finais.

2- Enquadramento

O final do século passado viu surgir um novo domínio de investigação nascido do cruzamento improvável da sintaxe generativa com a dialetologia. A emergência da sintaxe dialetal foi propiciada por três grandes conquistas teóricas e metodológicas: a modelização da faculdade da linguagem no âmbito do quadro de Princípios & Parâmetros (CHOMSKY, 1981), a formalização da teoria paramétrica do Programa Minimalista (CHOMSKY, 1995) e a promoção da abordagem microcomparativa no trabalho sobre variação sintática (KAYNE, 1996). A conjugação destes fatores definiu um novo programa para a investigação da variação sintática, no qual a análise comparativa de dialetos de uma mesma língua se anunciava como uma ferramenta poderosa: comparar variedades com gramáticas muito próximas, reduzindo a interferência das variáveis em análise, prometia potenciar a descoberta de propriedades sintáticas parametricamente associadas e, consequentemente, identificar os parâmetros primitivos da componente sintática da gramática e os princípios universais a eles subjacentes (KAYNE, 1996, 2005; POLETTO; BENINCÀ, 2007).

A adoção desta abordagem desencadeou, sobretudo na Europa, a criação de múltiplos projetos nacionais de sintaxe dialetal, que, num primeiro momento, se dedicaram a compilar extensas coleções de dados dialetais, dando origem a mais de duas dezenas de corpora e bases de dados informatizados (cf. e.o. GLASER; BART (2000); KORTMANN (2002); BARBIERS (2006); FLEISCHER; LENZ; WEISS (s.d.); LINDSTRÖM (s.d.); BENINCÀ (s.d.); VANGSNES (s.d.) e também ZANUTTINI; HORN; WOOD (2010), para desenvolvimentos mais recentes fora da Europa). Esta rede crescente de recursos alimentou, nos últimos anos, a investigação em variação sintática, agregando uma comunidade de

investigadores – sintaticistas e tipologistas – comprometidos com o estudo de fenómenos dialetais tradicionalmente considerados marginais quer à teoria sintática, quer à classificação tipológica (BARBIERS; CORNIPS; KLEIJ, 2002; KORTMANN, 2004; BARBIERS et al., 2008; KAYNE, 2013).

Portugal, e particularmente o Centro de Linguística da Universidade de Lisboa, aderiu ao programa da sintaxe dialetal desde a sua génese, avançando, em 1999, com um projeto de investigação dedicado ao estudo da variação sintática dialetal em PE e constituindo, com esse propósito, o Corpus Dialetal para o Estudo da Sintaxe (CORDIAL-SIN) (MARTINS, 2000). Pelas suas características, o CORDIAL-SIN configura um recurso único no banco de ferramentas para a sintaxe dialetal (CARRILHO, 2010; CARRILHO; MAGRO, 2010; MAGRO, 2010a; CARRILHO; MAGRO, 2011a; CARRILHO, 2012): é um corpus geograficamente representativo dos dialetos do PE composto por excertos de discurso espontâneo selecionados a partir de gravações de inquéritos dialetais realizados em 42 localidades do território português; com uma extensão de cerca de 600000 palavras, o CORDIAL-SIN é anotado ao nível da palavra e da frase seguindo protocolos de anotação morfossintática e sintática partilhados por diversos projetos nacionais e internacionais (SANTORINI, 2016; MAGRO; GALVES; CARRILHO, 2018), o que lhe garante interoperabilidade com outros corpora e bases de dados históricos (CLUL, 2014; MARTINS, 2015; GALVES; ANDRADE; FARIA, 2017) e dialetais (KUNST; WESSELING, 2011; BARBIERS, 2012).

A criação do CORDIAL-SIN fomentou a investigação sobre a sintaxe dos dialetos do PE, levando à identificação e estudo de fenómenos sintáticos desconhecidos e/ou não considerados na literatura sobre o português, como construções de duplo sujeito (MARTINS, 2009), construções de duplo objeto (CARDOSO; MAGRO, 2012), interpolação e duplicação de clítico (MAGRO, 2007a), expletivos periféricos (CARRILHO, 2005; BETONI, 2013; CARRILHO, 2014), relativas clivadas (CARDOSO; ALEXANDRE, 2013), clivadas de *é que* recursivo (COSTA; LOBO, 2009; VERCAUTEREN, 2010; VERCAUTEREN, 2015), gerúndio flexionado (LOBO, 2008, 2016) e outras manifestações de concordância não-padrão (COSTA; MOURA; PEREIRA, 2001; COSTA; PEREIRA, 2005; CARDOSO; CARRILHO; PEREIRA, 2011; COSTA; PEREIRA, 2013; SÓRIA, 2013). Os resultados destes trabalhos não só ampliaram o conhecimento sobre a gramática do português, como contribuíram, com nova evidência empírica, para o debate sobre teoria sintática. Assim, por exemplo, a inspeção

de dados dialetais trouxe novos argumentos a favor (i) da conexão entre computação sintática e certas propriedades discursivas, como exaustividade e força ilocutória (CARRILHO, 2005; BETONI, 2013; CARDOSO; ALEXANDRE, 2013; CARRILHO, 2014), (ii) da organização hierárquica dos traços-φ pronominais, em particular, do estatuto dissociado do traço de número (COSTA; PEREIRA, 2005; MARTINS, 2009; CARDOSO; CARRILHO; PEREIRA, 2011; COSTA; PEREIRA, 2013) ou (iii) de um modelo de organização da gramática que prevê a existência de movimento pós-sintático (MAGRO, 2007a) e inserção lexical tardia (COSTA; PEREIRA, 2005; COSTA; PEREIRA, 2013). Adicionalmente, a comparação entre as variedades dialetais representadas no CORDIAL-SIN e variedades históricas do português ou variedades dialetais de línguas geneticamente relacionadas, como o Português do Brasil ou o Galego, permitiu repensar as habituais lógicas de relação entre variação sincrónica e diacrónica, ora por se atribuir aos dialetos um papel inovador em certos processos de mudança (MAGRO, 2010b), ora por a diferenciação de variedades contemporâneas com um passado gramatical comum infirmarem correlações diacrónicas entre fenómenos (MAGRO, 2007a; SÓRIA, 2013).

Chegou agora o momento – já anunciado em trabalhos exploratórios (CARRILHO; PEREIRA, 2011; CARRILHO; MAGRO; ÁLVAREZ, 2013; CARRILHO; PEREIRA, 2013; PEREIRA, em preparação) – de considerar a variação sintática dialetal em PE a partir de uma perspetiva geolinguística, analisando a distribuição geográfica de fenómenos sintáticos no território português e identificando áreas de convergência linguística. Este passo, é essencial para responder aos desafios da teoria paramétrica e, assim, cumprir plenamente o programa da sintaxe dialetal: isolar variáveis sintáticas, conhecer a distribuição areal das respetivas variantes e observar as correlações entre variantes no espaço de uma mesma área geográfica será a estratégia para identificar os (micro)parâmetros que configuram diferentes sistemas gramaticais.

Dar este salto implica modernizar tecnologicamente o CORDIAL-SIN, transformando-o num recurso que melhor sirva esse objetivo. Este é o propósito do projeto SynAPse – Atlas Sintático do Português Europeu (MAGRO, em preparação), que se propõe cruzar avanços recentes nos domínios da mineração de florestas sintáticas (JANSSEN, 2016; MAGRO, 2018) e da representação cartográfica (LAMELI; KEHREIN; RABANUS, 2010) para criar uma ferramenta que mapeie automaticamente resultados de pesquisa de estruturas sintáticas. A construção de um atlas sintático dinâmico e interativo, como o que aqui

apresentamos, afigura-se não só uma solução inovadora em cartografia linguística, como a mais adequada ao estudo da dimensão espacial da sintaxe dialetal a partir de dados de corpora sintaticamente anotados (KUNST; BARBIERS, 2010; LAMELI; KEHREIN; RABANUS, 2010).

3- Objetivos e metodologia

Este projeto tem, pois, um duplo objetivo: (i) estudar a dimensão espacial da variação sintática em PE e (ii) construir um recurso digital que responda aos requisitos da investigação nesse domínio.

O objetivo enunciado em (i) tem uma importância central para o cumprimento do programa de investigação em sintaxe dialetal, que, em última instância, propõe trazer novos contributos para a teoria paramétrica através da análise comparada de variedades com gramáticas muito próximas. Este propósito tem de passar necessariamente pela delimitação prévia dessas variedades, o que assenta na análise da distribuição geográfica dos traços sintáticos não-padrão e na identificação de áreas de coesão sintática. Uma estratégia que combine o aparato teórico da sintaxe com a abordagem metodológica da geolinguística, nos termos da que se sugere no âmbito deste projeto, será aquela que serve esta finalidade. Assim, o objetivo (i) desdobra-se nos seguintes pontos: (a) identificação das construções sintáticas não-padrão (variáveis sintáticas) e respetivas variantes, (b) análise da distribuição areal no território português das variantes identificadas, (c) observação da correlação entre variantes no espaço de uma mesma área geográfica, (d) delimitação de áreas geográficas sintaticamente homogéneas.

Para além de criar condições para a reflexão sobre teoria paramétrica, esta abordagem permite paralelamente diferenciar a variação sintática dialetal de outros tipos de microvariação sintática (idioletal ou intrapessoal), que apresentam padrões de distribuição geográfica dispersa, contribuindo eventualmente para a construção de um modelo de variação que associa diferentes níveis de variação a diferentes módulos da gramática mental (JACKENDOFF, 2002; BARBIERS, 2013).

Adicionalmente, este projeto torna possível comparar de forma sistemática a distribuição geográfica de variantes sintáticas com a distribuição de variantes fonético-fonológicas, tradicionalmente consideradas na delimitação dos dialetos do PE

(CINTRA, 1971; CINTRA, 1990; SEGURA, 2006). A confirmarem-se os resultados dos primeiros estudos em geosintaxe do PE (CARRILHO; PEREIRA, 2011; CARRILHO; PEREIRA, 2013; PEREIRA, em preparação) – que apontam para a coincidência das fronteiras geográficas de áreas sintáticas e de áreas fonético-fonológicas –, sairão reforçadas a delimitação e diferenciação dos dialetos portugueses e ficará mais completa a sua caracterização empírica pela incorporação de nova evidência de base sintática.

Os resultados deste projeto trazem, pois, vantagens a dois domínios distintos: à teoria sintática (ou, em geral, à teoria da gramática), pela renovada perspetiva da análise microparamétrica, e à dialetologia, pela relevância do papel da sintaxe na formação do espaço linguístico.

Este programa de investigação assenta num outro objetivo: o da criação de um instrumento de observação da sintaxe no espaço. Neste plano, a opção pela construção de um atlas linguístico (neste caso, um atlas sintático) é a mais imediata. Desde logo porque os mapas são um meio privilegiado de visualização da disposição geográfica de factos linguísticos, mas sobretudo porque as representações cartográficas, mais do que o reflexo visual de uma análise, podem ser o próprio instrumento de análise, favorecendo a formulação de hipóteses sobre correlações espaciais.

Este é o caso do atlas a conceber, construir e disponibilizar no âmbito deste projeto. Com efeito, o Atlas Sintático do Português Europeu, pelo seu alto potencial heurístico, constituirá um instrumento (e não apenas um produto) da investigação em sintaxe dialetal. O desenho desta ferramenta é inspirado em ferramentas de referência da cartografia linguística moderna, como o *World Atlas of Language Structures Online* (DRYER; HASPELMATH, 2013) ou o *Dynamic Syntactic Atlas of Dutch Dialects* (BARBIERS, 2006), embora os desenvolvimentos agora propostos façam do SynAPse um recurso cartográfico de última geração. O seu caráter inovador resulta da conjugação de três características fundamentais: (i) ser um atlas digital dinâmico disponível em linha, (ii) gerar mapas definidos pelo utilizador e (iii) ser alimentado por um corpus sintaticamente anotado.

Um recurso com estes atributos não é simplesmente a contrapartida computorizada dos atlas linguísticos impressos – sempre resultados estáticos das escolhas e objetivos dos seus autores –, mas sim um verdadeiro laboratório de pesquisa, que oferece aos seus utilizadores: acesso em linha, modos dinâmicos de representação cartográfica, mapeamento automático de resultados de pesquisas sintáticas, combinação num mesmo mapa de resultados de pesquisas

múltiplas, geração automática de mapas síntese, traçado automático de isoglossas, quantificação automática de resultados, consulta dos dados sintáticos subjacentes aos mapas.

Esta ferramenta articula um corpus dialetal sintaticamente anotado (CORDIAL-SIN), um motor de busca para estruturas sintáticas (TEITOK-QUETch) e uma aplicação de webGIS (QGIS). A sua construção rentabiliza recursos computacionais já existentes no CLUL, nomeadamente o TEITOK, uma plataforma em linha, disponível em acesso aberto, desenvolvida por Maarten Janssen em 2014 para criar, editar e pesquisar corpora com marcação filológica e anotação linguística (JANSSEN, 2014).

O alojamento do CORDIAL-SIN no TEITOK, após a codificação dos dados em formato XML-TEI, permite a visualização e a pesquisa integral do corpus (cf. secção 4.). Para além da pesquisa por palavra, lema e etiqueta POS, os dados serão pesquisáveis por construção sintática através de expressões de busca em linguagem XPath, a linguagem comum para pesquisar nós numa hierarquia XML. As expressões de busca poderão ser compostas pelo utilizador na interface de pesquisa sintática do TEITOK ou por ele selecionadas de entre o extenso conjunto de expressões predefinidas do QUETch (MAGRO, 2018). A proveniência geográfica dos dados será codificada através do sistema de coordenadas x-y, o que permitirá restringir o escopo das pesquisas a uma determinada localidade ou área geográfica e mapear automaticamente os resultados na aplicação cartográfica articulada com o TEITOK. Uma vez que será possível combinar no mesmo mapa resultados de pesquisas múltiplas, o SynAPse será de extrema utilidade para a visualização de potenciais correlações entre fenómenos.

A análise linguística dos dados será desenvolvida numa perspetiva microcomparativa, contrastando-se o comportamento sintático das diferentes variedades dialetais identificadas (incluindo a variedade padrão). Uma abordagem comparativa teoricamente orientada (que no âmbito deste projeto seguirá a versão minimalista do quadro generativista de Princípios & Parâmetros) não só fornecerá os instrumentos de análise necessários à descrição e compreensão dos contrastes observados, como orientará a exploração do corpus, permitindo fazer predições e levantar novas questões (CINQUE; KAYNE, 2005).

A análise da distribuição geográfica dos dados terá como objetivo descrever a forma como os factos sintáticos se comportam no espaço, como se organizam e como se relacionam. A tarefa de identificar e delimitar áreas dialetais de base sintática e reconhecer padrões de distribuição geográfica de variantes sintáticas será abordada através dos métodos matemáticos

e estatísticos que suportam análise espacial, concretamente os métodos de reconhecimento de padrões de concentração/distribuição (e.g. medidas de interpolação espacial, como densidade de Kernel (KDE) e distância inversa ponderada (IDW), e medidas de autocorrelação espacial, como índice global de Moran e índice local Getis-Ord Gi*) (HOCH; HAYES, 2010).

Os resultados da análise linguística e espacial desenvolvida no âmbito do projeto reverterão também para a constituição do próprio SynAPse. Com efeito, este recurso, para além da funcionalidade de mapeamento automático e dinâmico dos dados do CORDIAL-SIN, disponibilizará igualmente uma seleção de mapas relativos a construções sintáticas não-standard que apresentem padrões de distribuição geolinguisticamente relevantes. A cada um destes mapas estará associado um texto informativo sobre o fenómeno representado, que incluirá uma breve descrição linguística da construção, um comentário ao padrão de distribuição geográfica que apresenta e uma bibliografia de referência.

4- Implementação

4.1- Codificação do corpus em XML-TEI

Devido à natureza espontânea do discurso que regista, à representatividade geográfica dos dados que integra e à anotação linguística que contém, o CORDIAL-SIN, como já se assinalou acima, constitui um recurso único para o estudo da sintaxe dialetal do PE. No plano linguístico, não existem, portanto, dúvidas quanto à sua utilidade. No entanto, no plano tecnológico, torna-se evidente que este corpus reflete a época em que foi construído e apresenta, consequentemente, algumas limitações importantes que dificultam, quando não impedem, uma exploração completa das suas potencialidades. Uma breve exposição do estado atual do CORDIAL-SIN permitirá compreender melhor essas limitações e, por conseguinte, as razões que levaram a converter o corpus em linguagem XML, o que corresponde ao objetivo inicial do projeto SynAPse.

Os materiais textuais do CORDIAL-SIN distribuem-se por quatro conjuntos de dados e estão armazenados em três formatos eletrónicos diferentes, conforme se mostra no Quadro 1. Todos os materiais estão disponíveis para descarga, com exceção das transcrições conservadoras em formato MS Word[ii]:

Tipo de dados	Tipo de formato
1. Transcrição conservadora	PDF, MS Word
2. Transcrição normalizada	PDF, Texto simples
3. Anotação morfossintática	Texto simples
4. Anotação sintática	Texto simples

Quadro 1- Distribuição e formato dos materiais do CORDIAL-SIN
Fonte: Elaboração própria

A transcrição conservadora contém a marcação de fenómenos típicos da oralidade, como a representação de variantes fonéticas e morfofonológicas ou a marcação de aspetos relacionados com a produção do discurso (abandono de fragmentos frásicos, repetições, formas truncadas, pausas, vocalizações, sobreposições de produção, etc.). A transcrição normalizada resulta da eliminação dos fenómenos de oralidade transcritos e codificados na versão conservadora e apresenta apenas a transcrição ortográfica dos dados expurgados. A versão com anotação morfossintática inclui etiquetas morfossintáticas para cada forma lexical da versão normalizada. O quarto e último conjunto de dados toma como ponto de partida a anotação morfossintática e apresenta a análise sintática de cada frase sob o formato de parentetização etiquetada.

A apresentação dos materiais do CORDIAL-SIN, exposta no Quadro 1, evidencia pelo menos dois problemas relevantes. O primeiro é relativo à preservação dos dados, uma questão que levanta especial preocupação no âmbito das humanidades digitais (SMITH, 2004). É sabido que o uso de formatos proprietários, por oposição a formatos de código aberto, constitui um risco para a conservação da informação digital, especialmente em situações de conflito entre interesses comerciais e preservação de dados:

> File formats that are proprietary are often identified as being especially at risk, because they are in principle dependent on support from an enterprise that may go out of business. Even a format so widely used that it is a *de facto* standard, such as Adobe Systems, Inc.'s portable document format (PDF), is treated with great caution by those responsible for persistence. The owner of a such a *de facto* standard has no legal obligation to release its source code or any other proprietary information in the event that it goes bankrupt or decides to stop supporting the format for one reason or another (such as creating a better and more lucrative file format) (SMITH, 2004, em linha).

Para além disso, o uso de software proprietário para o armazenamento de dados é particularmente problemático no caso da linguística de corpus, cuja metodologia implica o uso de ferramentas de análise que não são habitualmente compatíveis com formatos proprietários:

> If your corpus is made up of files in a format for a commercial word-processing program, such as Microsoft Word, then they cannot be processed by most corpus analysis tools. What is more, the format may not be supported indefinitely into the future, and there will come a time when users won't be able to read the files any more (WYNNE, 2005, em linha).

No caso do corpus CORDIAL-SIN, cujos materiais datam de 1999, o uso do Microsoft Word para armazenar os dados começava já a constituir um problema real de preservação no que respeita às transcrições conservadoras. Devido ao facto de a transcrição fonética usar uma fonte atualmente obsoleta (SIL DOULOS IPA93), foi necessário recorrer a ferramentas de conversão específicas para manter os caracteres fonéticos no âmbito do novo projeto.

O segundo problema relaciona-se com a recuperação da informação. É evidente que formatos de arquivo binários como PDF ou Microsoft Word não são apropriados para trabalhar em linguística de corpus, já que não foram concebidos para a recuperação de informação. Mas, para além disso, o corpus CORDIAL-SIN revela uma terceira limitação de não pouca importância. A distribuição dos materiais em quatro conjuntos de dados independentes dificulta a manutenção do corpus. Assim, a edição de uma ou várias palavras num conjunto de dados (por exemplo, a correção de um erro na transcrição conservadora) implica repetir manualmente a mesma correção nos outros três conjuntos de dados de modo a manter atualizado o corpus completo. Dito de outro modo, ter diferentes níveis de anotação de forma separada e sem vinculação interna significa manter quatro corpora independentes, com os custos em termos de tempo e de esforço que isso acarreta. Esta situação afeta igualmente as possibilidades de exploração do corpus porque impede a realização de pesquisas cruzadas entre os vários níveis de transcrição e/ou de anotação: não é possível, por exemplo, recuperar os pronomes clíticos (anotação morfológica) que apresentam realizações fonéticas particulares (transcrição conservadora), nem, por exemplo, extrair variantes morfológicas dialetais (transcrição conservadora) associadas a uma mesma forma padrão (transcrição normalizada) ou a uma determinada classe de palavras (anotação morfológica). O problema estende-se à

recuperação dos metadados, visto que a informação metatextual do CORDIAL-SIN está armazenada exclusivamente nos ficheiros Word (ou nas correspondentes versões em PDF) da transcrição conservadora, o que muito dificulta a realização de pesquisas cruzadas entre aspetos linguísticos e aspetos sociais.

Todas estas limitações são facilmente superáveis mediante a conversão do CORDIAL-SIN num corpus com formato standard que assegure a sua preservação a longo prazo, que facilite a sua compatibilização com outros recursos e ferramentas de análise e, por último, mas não menos importante, que melhore e amplie as possibilidades da sua exploração. Atualmente, esse formato é, sem dúvida, o XML:

> Open standards like XML are preferred because they make it possible to encode the intellectual content of the resource and the metadata in a consistent and unambiguous way. While there are reasons why XML, and Unicode, are desirable, and likely to become more firmly entrenched and widely used for language corpora, it is often trivial to migrate from other formats and standards, including proprietary ones, as long as good practice has been followed in the creation of the electronic text in whatever format (WYNNE, 2005, em linha).

> […] the goals of corpus linguistics and language documentation are not so different. Both fields aim for collections of related language data that are interoperable, searchable, reusable, and mobilizable for a broad range of linguistic inquiry […]. Current advances in encoding and interoperability like XML and Unicode are already making this possible (GRIES; BEREZ, 2017, p. 404).

Como assinala Wynne (2005) na citação acima, a conversão de outros formatos em linguagem XML não deve causar problemas significativos desde que se tenham adotado "boas práticas" no processo de criação dos dados de partida. Acima de tudo, é essencial que a anotação dos dados textuais tenha seguido critérios consistentes e sistemáticos, independentemente de esses critérios obedecerem ou não a padrões standardizados. É isto que se verifica no caso do CORDIAL-SIN, cujos critérios de anotação estão claramente explicitados e explicados, tanto para as transcrições do material sonoro (MAGRO, 2007b) como para as versões com anotação morfossintática (MAGRO; MORGADO, 2008) e sintática (CARRILHO; MAGRO, 2011b).

Para a conversão do CORDIAL-SIN num corpus em linguagem XML utilizaram-se apenas os materiais da transcrição conservadora e da anotação sintática, isto é, as versões 1 e 4 do Quadro 1. A transcrição normalizada está incluída nos materiais da transcrição

conservadora e, paralelamente, a anotação morfossintática está incluída nas versões com anotação sintática; ou seja, as versões 2 e 3 podem ser ignoradas no processo de conversão, uma vez que o material que contêm é facilmente extraível a partir das versões 1 e 2. A conversão foi realizada de forma automática através da aplicação sequencial de scripts baseados em linguagem Perl. Esquematicamente, o processo de codificação do corpus CORDIAL-SIN em linguagem XML pode resumir-se nos seguintes passos:

1. Conversão dos 42 documentos com a transcrição conservadora em formato Microsoft Word (.doc) em 42 documentos em texto simples Unicode (.txt).
2. Criação de um arquivo XML para cada sequência de inquérito do CORDIAL-SIN. O número total de arquivos XML que constituem o corpus é de 2058. Cada arquivo XML está dividido em metadados (<teiHeader>) e texto (<text>).
3. Conversão dos metadados.
4. Tokenização do texto.
5. Conversão das marcas textuais de transcrição.
6. Importação da anotação morfossintática.

Em conformidade com as práticas atuais no campo das humanidades digitais, neste processo de conversão do corpus foram adotadas as diretrizes de codificação propostas pelo consórcio TEI (Text Encoding Initiative) para a edição de textos em formato digital (TEI CONSORTIUM, 2019). Os standards TEI foram fundamentalmente aplicados à informação metatextual de cada ficheiro XML, isto é, à informação incluída no cabeçalho (<teiHeader>). Para a marcação de aspetos textuais, incluindo a própria tokenização do texto e a sua anotação linguística, adotou-se a estratégia de marcação utilizada pela ferramenta TEITOK (JANSSEN, 2014, 2016). O TEITOK é uma plataforma web especialmente desenhada para ver, criar e editar corpora que combinam marcação textual e anotação linguística, que é utilizada pelo projeto SynAPse para visualizar, editar e explorar os dados do CORDIAL-SIN uma vez codificados em XML. O sistema de visualização e processamento do TEITOK requer uma marcação que difere ligeiramente da aconselhada pelo TEI. No entanto, a marcação TEITOK é facilmente convertível em TEI standard de forma automática.

A título ilustrativo, apresentamos um breve fragmento textual do CORDIAL-SIN na sua versão conservadora (Figura 1) e, de seguida, o mesmo fragmento na versão do SynAPse,

ou seja, o resultado da aplicação da sequência de scripts que fazem a conversão em XML e a importação da anotação morfossintática (Figura 2). A versão em XML aqui apresentada cinge-se ao fragmento produzido pelo informante e está ligeiramente simplificada para clareza da exposição [iii].

INQ1 E antigamente, para acartar pedra, não havia umas coisas que se arrastavam pelo chão?

INF1 Ah, isso são os arrastões. [AB|Isso é] Isso é assim. [AB|Isso é à maneira dum] Isso é um arrastão. É um arrastão… Quer dizer, é assim neste processo. Olhe. {pp} Compreende? {fp} {PH|suˈpoɲɐmuz=Suponhamos} isto.

Figura 1- Fragmento de texto no CORDIAL-SIN (versão conservadora)
Fonte: Elaboração própria

```
<u who="#INF1">
        <tok id="w-19">INF1</tok>
        <tok pos="INTJ" id="w-20">Ah</tok>
        <tok pos="," id="w-21">,</tok>
        <tok pos="DEM" id="w-22">isso</tok>
        <tok pos="SR-P-3P" id="w-23">são</tok>
        <tok pos="D-P" id="w-24">os</tok>
        <tok pos="N-P" id="w-25">arrastões</tok>
        <tok pos="." id="w-26">.</tok>
        <seg type="abandoned" >
                <tok inform="Isso" nform="--" id="w-27">Isso</tok>
                <tok inform="é" nform="--" id="w-28">é</tok>
        </seg>
                <tok pos="DEM" id="w-29">Isso</tok>
                <tok pos="SR-P-3S" id="w-30">é</tok>
                <tok pos="ADV" id="w-31">assim</tok>
                <tok pos="." id="w-32">.</tok>
        <seg type="abandoned" >
                <tok inform="Isso" nform="--" id="w-33">Isso</tok>
                <tok inform="é" nform="--" id="w-34">é</tok>
                <tok inform="à" nform="--" id="w-35">à</tok>
                <tok inform="maneira" nform="--" id="w-36">maneira</tok>
                <tok inform="dum" nform="--" id="w-37">dum</tok>
        </seg>
        <tok pos="DEM" id="w-38">Isso</tok>
        <tok pos="SR-P-3S" id="w-39">é</tok>
        <tok pos="D-UM" id="w-40">um</tok>
        <tok pos="N" id="w-41">arrastão</tok>
        <tok pos="." id="w-42">.</tok>
        <tok pos="SR-P-3S" id="w-43">É</tok>
        <tok pos="D-UM" id="w-44">um</tok>
        <tok pos="N" id="w-45">arrastão</tok>
        <tok pos="." id="w-46">…</tok>
        <tok pos="VB-P-3S" id="w-47">Quer</tok>
        <tok pos="VB" id="w-48">dizer</tok>
        <tok pos="," id="w-49">,</tok>
        <tok pos="SR-P-3S" id="w-50">é</tok>
        <tok pos="ADV" id="w-51">assim</tok>
        <tok id="w-52">neste
                <dtok form="em" pos="P" id="d-52-1"/>
                <dtok form="este" pos="D" id="d-52-2"/>
        </tok>
        <tok pos="N" id="w-53">processo</tok>
        <tok pos="." id="w-54">.</tok>
        <tok pos="VB-SP-3S" id="w-55">Olhe</tok>
        <tok pos="." id="w-56">.</tok>
        <pause>
                <tok psform="[pausa]" nform="--" id="w-57">[pausa]</tok>
        </pause>
        <tok pos="VB-P-3S" id="w-58">Compreende</tok>
        <tok pos="." id="w-59">?</tok>
        <vocal>
                <tok psform="[vocalização]" nform="--" id="w-60">[vocalização]</tok>
        </vocal>
        <tok phform="su ˈpoɲɐmuz" ipa="su ˈpoɲɐmuz" pos="VB-SP-1P" id="w-61">Suponhamos</tok>
        <tok pos="DEM" id="w-62">isto</tok>
        <tok pos="." id="w-63">.</tok>
</u>
```

Figura 2- Fragmento de texto codificado no SynAPse (XML)
Fonte: Elaboração própria

Nos documentos XML processados pelo TEITOK, cada token está marcado com um elemento <tok>, concetualmente semelhante ao elemento <w> proposto pelo TEI, mas com a particularidade de poder ser usado para qualquer tipo de anotação linguística. Na verdade, a diferença fundamental entre o TEITOK e o TEI assenta no conjunto de atributos do elemento <tok>. Como se vê na Figura 2, cada token (isto é, cada elemento <tok>) está associado a um identificador único através do atributo @id e a uma etiqueta morfossintática através do atributo @pos:

```
<tok pos="SR-P-3P" id="w-23">são</tok>
```

Nos casos em que uma única palavra ortográfica corresponde a duas ou mais palavras gramaticais, é usado o elemento <dtok/> dentro do elemento <tok> para poder associar a informação linguística a cada um dos formantes da palavra ortográfica (principalmente, casos de contrações e enclíticos):

```
<tok id="w-52">neste
    <dtok form="em" pos="P" id="d-52-1"/>
    <dtok form="este" pos="D" id="d-52-2"/>
</tok>
```

Outras informações relevantes relativas a cada token são expressas através de atributos específicos (ver Quadro 3, abaixo). Assim, por exemplo, as variantes fonéticas, que no CORDIAL-SIN são marcadas pelo código {PH}, são associadas no SynAPse a um atributo @phform:

```
<tok phform="su ˈ poɲɐmuz" ipa="su ˈ poɲɐmuz" pos="VB-SP-1P" id="w-61">Suponhamos</tok>
```

O mesmo se aplica às sequências abandonadas. O código [AB] utilizado no CORDIAL-SIN é substituído no SynAPse pelo atributo @inform (de incident form). Tais fragmentos frásicos (abandonados em resultado de processos de reformulação, adiamento da produção ou hesitação) não são visualizados na versão com texto normalizado, o que, por sua vez, se consegue no TEITOK aplicando dois travessões (--) ao atributo associado a essa

visualização, no caso @nform:

```
<tok inform="Isso" nform="--" id="w-33">Isso</tok>
```

Finalmente, o SynAPse utiliza elementos TEI para marcar dentro do texto determinadas informações que incluem um ou mais tokens. Por exemplo, os enunciados são delimitadas pelo elemento <u> (utterance), cujo atributo @who permite identificar o responsável pelo discurso. Outros exemplos de marcação de nível superior a token são as pausas (<pause>), as vocalizações (<vocal>) ou a informação extralinguística (<kinesic>)[iv]. Apresentamos, no Quadro 2, a lista completa de correspondências entre a marcação do CORDIAL-SIN e a marcação em XML do SynAPse, e, no Quadro 3, o inventário e a descrição sumária dos atributos de <tok> utilizados. Para clareza da exposição, os dados XML do Quadro 2 foram, também neste caso, simplificados:

Descrição	**CORDIAL-SIN**	**SynAPse/TEITOK**
Variantes fonéticas	{PH| nũ=não}	<tok phform="nũ">não</tok>
Formas contraídas	{CT| pa∫=para as}	<tok ctform="pa∫">para as</tok>
Formas truncadas	{IP| ˈtivɨ=estive}	<tok ipform="ˈtivɨ" ipa="ˈtivɨ">estive</tok>
Palavras estrangeiras	{FR| si ˈnɔ='snow'}	<foreign> <tok dform="sinó" ipa="si ˈnɔ">'snow'</tok> </foreign>
Variantes morfológicas	'ouvisto'	<tok dform="ouvisto" stform="ouvido">ouvisto</tok>
Itens lexicais não dicionarizados	'lúbis'	<tok dform="lúbis" stform="lobisomem">lúbis</tok>
Pausas não vocalizadas	{pp}	<pause> <tok psform="[pausa]">[pausa]</tok> </pause>
Pausas vocalizadas	{fp}	<vocal> <tok psform="[vocalização]">[vocalização]</tok> </vocal>
Sobreposições de produção	Aquí	<seg type="overlap_beg "/> <tok ovform="Aqui">Aqui</tok> <seg type="overlap_end "/>
Sequências abandonadas	[AB|Isso]	<seg type="abandoned"/> <tok inform="Isso">Isso</tok> </seg>
Sequências repetidas	[RP|sobre]	<seg type="repeated"/> <tok inform="sobre">sobre</tok> </seg>
Dúvidas	(ali)	<supplied cert="medium">

de audição		<tok>ali</tok> </supplied>
Divergências de audição	(e) /em\	<choice> <supplied cert="fair"><tok>e</tok></supplied> <supplied cert="low"><tok>em</tok></supplied> </choice>
Sequências impercetíveis	(…)	<tok> <gap reason="inaudible">[…]</gap> </tok>
Sequências impercetíveis com anotação POS	(…/N)	<supplied cert="high"> <tok pos="N">[N]</tok> </supplied>
Formas inacabadas	{RC| casa-=casado}	<supplied cert="high" reason="reconstructed"> <tok brform="casa-" nform="casado">casa-</tok> </supplied>
Informação extralinguística	[Risos]	<kinesic> <tok>Risos</tok> </kinesic>

Quadro 2- Correspondência entre a marcação do CORDIAL-SIN e do SynAPse/TEITOK
Fonte: Elaboração própria

Descrição	**Atributos**
Identificador	@id
Etiqueta POS	@pos
Forma normalizada	@nform
Variante fonética	@phform
Forma contraída	@ctform
Forma truncada	@ipform
Forma dialetal	@dform
Forma padrão	@stform
Forma sobreposta	@ovform
Forma abandonada / repetida	@inform
Forma inacabada	@brform
Transcrição fonética	@ipa
Pausa (vocalizada ou não)	@psform

Quadro 3- Lista de atributos de <tok> utilizados no SynAPse
Fonte: Elaboração própria

As vantagens que o TEITOK oferece são múltiplas. Em primeiro lugar, simplifica-se o processo de edição e manutenção do corpus, já que apenas é necessário atualizar uma fonte de dados, evitando-se assim os problemas causados pela existência de múltiplas versões independentes de um mesmo documento. Em segundo lugar, melhora-se a visualização dos dados, visto que uma interface baseada em HTML permite ao utilizador selecionar através de simples botões a versão do texto que considera mais adequada aos seus objetivos de consulta (conservadora, normalizada, anotada, etc.). E, em terceiro lugar, ampliam-se as possibilidades de exploração do corpus, criado automaticamente a partir do conjunto de documentos XML e integralmente pesquisável através da linguagem CQP (CHRIST et al., 1999) a partir da própria plataforma web. Uma vez que toda a informação dos documentos XML é exportável para o corpus CQP, é possível realizar pesquisas cruzadas entre aspetos textuais e metatextuais ou entre os diferentes níveis de informação associados a cada token.

4.2- Tarefas futuras

Na sequência das operações descritas na secção anterior, está atualmente alinhada, num mesmo documento XML, informação que, na versão original do CORDIAL-SIN, se encontrava dispersa em documentos vários: para cada sequência de inquérito dialetal incluída no corpus, existe agora um objeto digital único que integra metadados, transcrição dos dados orais, marcação associada aos diferentes níveis de transcrição e anotação morfossintática.

Ocupa-nos presentemente a codificação da anotação sintática e a sua vinculação aos restantes dados do corpus. Para tal, adotamos a solução de anotação em *stand-off* já testada no âmbito do projeto P.S. Post Scriptum (CLUL, 2014; JANSSEN, 2016): a anotação sintática é codificada em XML e interligada com o restante material por um processo de coindexação.

Num futuro mais distante, dedicar-nos-emos à construção da aplicação cartográfica, tarefa que, previsivelmente, envolverá: (i) a preparação da interface de pesquisa sintática do TEITOK para armazenamento em tabelas GEOxml de resultados georeferenciados de buscas múltiplas e sequenciais; (ii) o desenvolvimento de scripts em Perl para conversão do output de pesquisa do TEITOK em diferentes ficheiros CSV, correspondentes às diferentes camadas de um mapa; (iii) a criação de um projeto local com o software SIG QGIS para desenho do template cartográfico base; (iv) a conversão do projeto local num projeto em linha com o

plugin QGIS2WEB; (v) a exportação dos dados para a biblioteca em JavaScript OpenLayers para desenvolvimento das funções interativas.

5- Conclusão

Vimos neste artigo que a metodologia adotada no projeto SynAPse serve o duplo propósito de documentar e compreender a dimensão espacial da variação sintática em PE. O Atlas Sintático do Português Europeu responde a estes objetivos permitindo aos utilizadores visualizar os dados, interrogá-los e obter condições favoráveis à sua correta interpretação. Para quem estuda a diversidade linguística no espaço do território português, este recurso será, assim, uma poderosa ferramenta de trabalho, potencialmente reveladora de novas linhas de investigação em sintaxe dialetal e transformadora do conhecimento produzido nesta área. O que está aqui em causa não é, pois, um projeto que convoca práticas tecnológicas para pensar as humanidades; é a recriação das próprias humanidades através da tecnologia.

6- Referências

BARBIERS, S. **Dynamic Syntactic Atlas of the Dutch Dialects (Dynasand)**. Amsterdam: Meertens Institute, 2006. Disponível em: <http://www.meertens.knaw.nl/sand/>.

_____. **European Dialect Syntax (Edisyn)**. Amsterdam: Meertens Institute, 2012. Disponível em: <http://www.dialectsyntax.org/wiki/Main_Page>.

_____. Where is syntactic variation? In: AUER, P.; REINA, J. C.; KAUFMANN, G. (ed.). **Language Variation – European Perspectives IV**. Amsterdam/Philadelphia: John Benjamins, 2013. p. 1-26.

BARBIERS, S.; CORNIPS, L.; KLEIJ, S. V. D. (ed.). **Syntactic Microvariation**. Amsterdam: Meertens Institute, 2002.

BARBIERS, S. et al. (ed.). **Microvariation in syntactic doubling**. Chicago: Brill, 2008.

BENINCÀ, P. **Syntactic Atlas of Italy (Asit)**. Padova / Venezia: Università di Padova / Università di Venezia, s.d. Disponível em: <http://svrims2.dei.unipd.it:8080/asit-enterprise/>.

BETONI, S. **O expletivo ele em domínios dependentes em Português Europeu**. 2013. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade de Lisboa, Lisboa.

CARDOSO, A.; ALEXANDRE, N. Relativas clivadas em variedades não standard do português europeu. In: SILVA, F.; FALÉ, I.; PEREIRA, I. (ed.). **Textos Selecionados do XVIII Encontro Nacional da Associação Portuguesa de Linguística**. Porto: Associação Portuguesa de Linguística, 2013. p. 205-227.

CARDOSO, A.; CARRILHO, E.; PEREIRA, S. On verbal agreement variation in european portuguese: syntactic conditions for the 3SG/3PL alternation. **Diacrítica**, Braga, v. 25, n.1, p. 137-160, 2011.

CARDOSO, A.; MAGRO, C. The syntax of naming constructions in European Portuguese dialects: variation and change. **Journal of Portuguese Linguistics**, Lisboa, v. 11, n.1, p. 23-43, 2012.

CARRILHO, E. **Expletive ele in European Portuguese dialects.** 2005. Dissertação (Doutoramento em Linguística) – Faculdade de Letras, Universidade de Lisboa,

_____. Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects). In: AURREKOETXEA, G.; ORMAETXEA, J. L. (ed.). **Tools for Linguistic Variation**. Bilbao: Universidad del Pais Vasco, 2010. p. 57-70.

_____. **Using an annotated corpus for dialect syntax: the case of European Portuguese in CORDIAL-SIN**. Comunicação apresentada em VIIth Congress of the International Society for Dialectology and Geolinguistics (SIDG), Wien, 2012.

_____. **Demonstrativos neutros não-argumentais no CORDIAL-SIN: novos aspectos da interface sintaxe-discurso na periferia esquerda da frase**. Comunicação apresentada em V Encuentro Wedisyn, Madrid, 2014.

CARRILHO, E.; MAGRO, C. A Anotação sintáctica do CORDIAL-SIN. In: BRITO, A. M. et al. (ed.). **Textos Seleccionados do XXV Encontro Nacional da Associação Portuguesa de Linguística**. Porto: Associação Portuguesa de Linguística, 2010. p. 225-241.

_____. **Syntactic Annotation for Dialect Syntax: CORDIAL-SIN — The annotated corpus of Portuguese Dialects**. Póster apresentado em 6th International Conference on Language Variation in Europe (ICLaVE), Freiburg, 2011a.

_____. **CORDIAL-SIN. Syntax-oriented Corpus of Portuguese Dialects. Syntactic Annotation System Manual**. Lisboa: Centro de Linguística, Universidade de Lisboa, 2011b. Disponível em: <http://www.clul.ulisboa.pt/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html>.

CARRILHO, E.; MAGRO, C.; ÁLVAREZ, X. (ed.). **Current Approaches to Limits and Areas in Dialectology**. Newcastle upon Tyne: Cambridge Scholars Publishing, 2013.

CARRILHO, E.; PEREIRA, S. Sobre a distribuição geográfica de construções sintácticas não-padrão em português europeu. In: COSTA, A.; BARBOSA, P.; FALÉ, I. (ed.). **Textos Seleccionados do XXVI Encontro da Associação Portuguesa de Linguística**. Lisboa: Associação Portuguesa de Linguística, 2011. p. 125-139.

_____. On the areal dimension of non-standard syntax: Evidence from a Portuguese corpus. In: BARYSEVICH, A.; D'ARCY, A.; HEAP, D. (ed.). **Proceedings of Methods XIV: Papers from the Fourteenth International Conference on Methods in Dialectology**. Pieterlen: Peter Lang, 2013. p. 69-79.

CHOMSKY, N. **Lectures on Government and Binding**. Dordrecht: Foris, 1981.

_____. **The Minimalist Program**. Cambridge, Massachusetts: MIT Press, 1995.

CHRIST, O. et al. **The Ims Corpus Workbench: Corpus Query Processor (Cqp): User's Manual**. Sttutgart: University of Sttutgart, 1999.

CINQUE, G.; KAYNE, R. (ed.). **The Oxford Handbook of Comparative Syntax**. New York: Oxford University Press, 2005.

CINTRA, L. F. L. Nova proposta de classificação dos dialectos galego-portugueses. **Boletim de Filologia**, Lisboa, v. 22, p. 81-116, 1971.

_____. Os dialectos da ilha da Madeira no quadro dos dialectos galego-portugueses. In: FRANCO, J. E. (ed.). **Cultura Madeirense. Temas e Problemas**. Porto: Campo das Letras, 1990. p. 95-107.

CLUL (Coord.). **P.S. Post Scriptum. A Digital Archive of Ordinary Writings (Early Modern Portugal and Spain).** Lisboa: Centro de Linguística, Universidade de Lisboa, 2014. Disponível em: <http://ps.clul.ul.pt>.

COSTA, J.; LOBO, M. Estruturas clivadas: evidência dos dados do Português Europeu não standard. In: 2009, João Pessoa. **Anais do Congresso Internacional da Abralin 2**. João Pessoa: Ideia, 2009. p. 3800-3806.

COSTA, J.; MOURA, D.; PEREIRA, S. Concordância com a gente: um Problema para a Teoria de Verificação de Traços. In: Actas do XVI Encontro Nacional da Associação Portuguesa de Linguística, 2001, Lisboa. Lisboa: Associação Portuguesa de Linguística, 2001. p. 637-657.

COSTA, J.; PEREIRA, S. Phases and autonomous features: a case of mixed agreement in European Portuguese. In: MCGINNIS, M.; RICHARDS, N. (ed.). **Perspectives on Phases**. Cambridge, Massachusetts: MIT Press, 2005. p. 115-124.

_____. 'a gente': pronominal status and agreement revisited. **The Linguistic Review**, Berlin/Boston, v. 30, n.2, p. 161-184, 2013.

DRYER, M.; HASPELMATH, M. **The World Atlas of Language Structures Online.** Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. Disponível em: <http://wals.info>.

FLEISCHER, J.; LENZ, A.; WEISS, H. **Syntax of Hessian Dialects (Syhd)**. Marburg/Wien/Frankfurt: Philipps-Universität Marburg/ Universität Wien/Goethe Universität, s.d. Disponível em: <http://www.syhd.info/en/home/>.

GALVES, C.; ANDRADE, A. L.; FARIA, P. **Tycho Brahe Parsed Corpus of Historical Portuguese**. Campinas: Instituto de Estudos da Linguagem, Universidade de Campinas, 2017. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.>.

GLASER, E.; BART, G. **Dialect Syntax of Swiss German**. Zürich: Universität Zürich, 2000. Disponível em: <http://www.research-projects.uzh.ch/p1794.htm>.

GRIES, S. T.; BEREZ, A. L. Annotation in/for Corpus Linguistics. In: IDE, N. P., James (ed.). **Handbook of Linguistic Annotation**. Berlin: Springer, 2017. p. 379-410.

HOCH, S.; HAYES, J. J. Geolinguistics: The incorporation of geographic information systems and science. **The Geographical Bulletin**, Lawrence Ville, Georgia, v. 51, p. 23-36, 2010.

JACKENDOFF, R. **Foundations of Language. Brain, Meaning, Grammar, Evolution.** Oxford: Oxford University Press, 2002.

JANSSEN, M. **TEITOK – The tokenized TEI environment**. Lisboa: Centro de Linguística da Universidade de Lisboa, 2014. Disponível em: <http://alfclul.clul.ul.pt/teitok/site/index.php?action=home>.

_____. TEITOK: Text-Faithful Annotated Corpora. In: Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, Portorož. Portorož: European Language Resources Association, 2016. p. 4037-4043.

KAYNE, R. Microparametric Syntax: Some Introductory Remarks. In: BLACK, J.; MOTAPANYANE, V. (ed.). **Microparametric Syntax and Dialect Variation**. Amsterdam/Philadelphia: John Benjamins, 1996. p. ix-xviii.

_____. Some notes on comparative syntax, with special reference to English and French. In: CINQUE, G.; KAYNE, R. (ed.). **The Oxford Handbook of Comparative Syntax**. New York: Oxford University Press, 2005. p. 3-69.

_____. Comparative syntax. **Lingua**, Amsterdam, v. 130, p. 132-151, 2013.

KORTMANN, B. **Freiburg English Dialect Corpus**. Freiburg: Universität Freiburg, 2002. Disponível em: <http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/>.

_____. **Dialectology Meets Typology: Dialect Grammar From a Cross-Linguistic Perspective.** Berlim: Walter de Gruyter, 2004.

KUNST, J. P.; BARBIERS, S. Generating maps on the internet. In: LAMELI, A.; KEHREIN, R.; RABANUS, S. (ed.). **Language and Space. An International Handbook of Linguistic Variation. Volume 2 Language Mapping**. Berlin: De Gruyter, 2010. p. 401-422.

KUNST, J. P.; WESSELING, F. The Edisyn search engine. **Oslo studies in language**, Oslo, v. 3, n.2, p. 63-74, 2011.

LAMELI, A.; KEHREIN, R.; RABANUS, S. Introduction. In: LAMELI, A.; KEHREIN, R.; RABANUS, S. (ed.). **Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping. Part I**. Berlin/New York: De Gruyter Mouton, 2010. p. xi-xxii.

LINDSTRÖM, L. **Estonian Dialect Corpus**. Tartu: University of Tartu / Institute of Estonian Language, s.d. Disponível em: <http://www.murre.ut.ee/estonian-dialect-corpus/>.

LOBO, M. Variação morfo-sintáctica em dialectos do Português europeu: o gerúndio flexionado. **Diacrítica**, Braga, v. 22, n.1, p. 25-55, 2008.

_____. O gerúndio flexionado no português dialetal. In: MARTINS, A. M.; CARRILHO, E. (ed.). **Manual de linguística portuguesa**. Berlin: Mouton de Gruyter, 2016. p. 481-501.

MAGRO, C. **Clíticos: Variações sobre o tema**. 2007a. Dissertação (Doutoramento em Linguística) – Faculdade de Letras, Universidade de Lisboa, Lisboa.

_____. (ed.) **CORDIAL-SIN. Corpus Dialectal para o Estudo da Sintaxe. Normas de Transcrição.** Lisboa: Centro de Linguística, Universidade de Lisboa, 2007b. Disponível em: <http://www.clul.ulisboa.pt/english/sectores/variacao/cordialsin/manual_normas.pdf>.

_____. When CORDIAL becames friendly: endowing the CORDIAL-SIN corpus with a syntactic annotation layer. In: International Conference on Language Resources and Evaluation, 7, 2010a, La Valetta. **Proceedings from Seventh International Conference on Language Resources and Evaluation (LREC 2010)**. La Valetta: European Language Resources Association, 2010a. p. 3705-3711.

_____. When corpus analysis refutes common beliefs. The case of interpolation in European Portuguese dialects. **Corpus**, Nice, v. 9, p. 115-135, 2010b.

_____. **QETch. Queries for Tree Searching**. Lisboa: Centro de Linguística da Universidade de Lisboa, 2018. Disponível em: <http://ps.clul.ul.pt/index.php?action=treequeries>.

_____. (coord.). **Synapse – Syntactic Atlas of European Portuguese / Atlas Sintático Do Português Europeu**. Lisboa: Centro de Linguística da Universidade de Lisboa, em preparação. Disponível em: <http://cards-fly.clul.ul.pt/teitok/synapse/>.

MAGRO, C.; GALVES, C.; CARRILHO, E. **Portuguese Syntactic Annotation Manual**. Lisboa/Campinas: Centro de Linguística, Universidade de Lisboa/Instituto de Estudos da Linguagem, Universidade de Campinas, 2018. Disponível em: <https://sites.google.com/site/portuguesesyntacticannotation/>.

MAGRO, C.; MORGADO, C. (ed.) **CORDIAL-SIN. Syntax-oriented Corpus of Portuguese Dialects. POS Annotation Manual**. Lisboa: Centro de Linguística, Universidade de Lisboa, 2008. Disponível em: <http://www.clul.ulisboa.pt/english/sectores/variacao/cordialsin/pos_annotation_manual.pdf.

MARTINS, A. M. (Coord.). **Corpus Dialetal Para O Estudo Da Sintaxe / the Syntax-Oriented Corpus of Portuguese Dialects (Cordial-Sin)**. Lisboa: Centro de Linguística da Universidade de Lisboa, 2000. Disponível em: <http://www.clul.ulisboa.pt/en/11-resources/314-cordial-sin-corpus-2>.

_____. Subject doubling in European Portuguese dialects: The role of impersonal se. In: ABOH, E. O. et al. (ed.). **Romance Languages and Linguistic Theory 2007**. Amsterdam/Philadelphia: John Benjamins, 2009. p. 179-200.

_____. (Coord.). **Word Order and Word Order Change in Western European Languages (Wochwel)**. Lisboa: Centro de Linguística da Universidade de Lisboa, 2015. Disponível em: <http://alfclul.clul.ul.pt/wochwel/>.

PEREIRA, S. A. **Áreas sintáticas no território português**. em preparação. Dissertação (Doutoramento em Linguística) – Faculdade de Letras, Universidade de Lisboa, Lisboa.

POLETTO, C.; BENINCÀ, P. The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects. **Nordlyd**, Tromsø, v. 34, n.1, p. 35-52, 2007.

SANTORINI, B. **Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence**. Philadelphia: Department of Linguistics, University of Pennsylvania, 2016. Disponível em: <https://www.ling.upenn.edu/hist-corpora/annotation/index.html>.

SEGURA, L. Dialectos açorianos. Contributos para a sua classificação. In: I Encontro de Estudos Dialectológicos – Actas, 2006, Ponta Delgada. Ponta Delgada: Instituto Cultural de Ponta Delgada, 2006. p. 325-344.

SMITH, A. Preservation. In: SCHREIBMAN, S.; SIEMENS, R.; UNSWORTH, J. (ed.). **A Companion to Digital Humanities**. Oxford: Blackwell, 2004. Disponível em: <http://www.digitalhumanities.org/companion/>

SÓRIA, M. **'Nós', 'a gente' e o sujeito nulo de primeira pessoa do plural**. 2013. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Universidade de Lisboa,

TEICONSORTIUM (ed.) **TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.5.0. Last updated on 29th January 2019.** TEI Consortium, 2019. Disponível em: <http://www.tei-c.org/Guidelines/P5/.>.

VANGSNES, Ø. A. **Scandiasyn – Scandinavian Dialect Syntax**. Tromsø: University of Tromsø, s.d. Disponível em: <http://websim.arkivert.uit.no/scandiasyn/index.html%3fLanguage=en>.

VERCAUTEREN, A. **Como é que é com o 'é que'? Análise de estruturas com 'é que' em variedades não-standard do Português Europeu**. 2010. Dissertação (Mestrado em Linguística) – Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa.

_____. **A conspiracy theory for clefts: the syntax and interpretation of cleft constructions**. 2015. Dissertação (Doutoramento em Linguística) – Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa.

WYNNE, M. Archiving, Distribution and Preservation. In: WYNNE, M. (ed.). **Developing Linguistic Corpora: A Guide to Good Practice**. Oxford: Oxford Books, 2005. Disponível em: <https://ota.ox.ac.uk/documents/creating/dlc/>

ZANUTTINI, R.; HORN, L.; WOOD, J. **Yale Grammatical Diversity Project: English in North America**. New Haven, Connecticut: Yale University, 2010. Disponível em: <http://ygdp.yale.edu>.

Sobre os autores

Catarina Magro. Doutorada em Linguística pela Universidade de Lisboa. Investigadora Auxiliar do Centro de Linguística da Universidade de Lisboa (Grupo de Investigação de Dialetologia & Diacronia). Investigadora Responsável do projeto SynAPse.

Gael Vaamonde. Doutorado em Linguística pela Universidade de Vigo. Professor Auxiliar na Faculdade de Filosofia e Letras da Universidade de Granada (Departamento de Língua Espanhola). Investigador Corresponsável do projeto SynAPse.

Notas

[i] O projeto SynAPse – Syntactic Atlas of European Portuguese / Atlas Sintático do Português Europeu é financiado pela Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior (PTDC/LLT-LIN/32086/2017).

[ii] Os materiais do CORDIAL-SIN estão disponíveis para descarga através do seguinte URL: <http://www.clul.ulisboa.pt/en/10-research/314-cordial-sin-corpus>

[iii] O documento XML completo que serviu de fonte a este exemplo pode ser descarregado a partir do seguinte endereço: <http://cards-fly.clul.ul.pt/teitok/synapse/index.php?action=file&id=AAL/AAL50.xml>

[iv] Estes elementos são conformes às diretrizes TEI, módulo 8, consultável no seguinte URL: <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>