A anotação sintáctica do CORDIAL-SIN

Ernestina Carrilho[#] & *Catarina Magro*
[#] Faculdade de Letras da Universidade de Lisboa
Centro de Linguística da Universidade de Lisboa

Abstract

The annotated dialect corpus CORDIAL-SIN provides straightforward and systematic access to a compilation of spontaneous and semi-directed speech from European Portuguese dialects. Optimal inquiry for precise morphological and syntactic information throughout this corpus is granted by two layers of annotation (part-of-speech tagging and syntactic annotation), which are automatically searchable for linguistic labels and structures through independently available tools. This paper concentrates on issues pertaining to the CORDIAL-SIN syntactic annotation: general aims, parsing procedures, the annotation system. Ultimately, the text casts light on the potential of the CORDIAL-SIN corpus as a tool for (dialect) syntactic inquiry.

Keywords: annotated corpus, syntactic annotation, dialect syntax, dialect corpus, European Portuguese dialects.
Palavras-chave: corpus anotado, anotação sintáctica, sintaxe dialectal, corpus dialectal, dialectos do português europeu.

1. CORDIAL-SIN: o *corpus* dialectal português para o estudo da sintaxe

O CORDIAL-SIN (*CORpus DIALectal para o estudo da SINtaxe*) integra textos orais provenientes de 42 localidades ou micro-regiões do território português, numa extensão de 600 000 palavras. Este *corpus* foi constituído no âmbito de diversos projectos de investigação desenvolvidos no Centro de Linguística da Universidade de Lisboa (CLUL)[1], comprometidos com o objectivo de impulsionar a investigação em sintaxe dia-

[1] Projectos financiados pela *Fundação para a Ciência e a Tecnologia* (PRAXIS XXI/P/PLP/13046/1998; POSI/1999/PLP/33275; POCTI/LIN/46980/2002; e, actualmente em curso, PTDC/LIN/71559/2006).

***Textos Seleccionados, XXV Encontro Nacional da Associação Portuguesa de Linguística,* Porto, APL, 2010, pp. 225-241**

lectal no domínio do português e de criar e disponibilizar um recurso linguístico capaz de satisfazer os requisitos empíricos específicos desta área de estudos[2].

Os textos compilados no CORDIAL-SIN correspondem a excertos de discurso espontâneo e semi-dirigido, de inquéritos dialectais realizados de forma sistemática no território português continental e insular para diferentes projectos de geografia linguística (ALEAç, ALEPG, ALLP e BA), integrantes do Arquivo Sonoro do CLUL. A procedência dos textos assim reunidos assegura a uniformidade do *corpus* relativamente à caracterização das localidades representadas (todas rurais e de pequenas dimensões), às condições da recolha linguística (situação de inquérito, equipa de inquiridores) e ao perfil dos informantes. Estes inquéritos dialectais do Arquivo Sonoro do CLUL, pela sua duração (de vários dias), pelo tipo de questionário por que se orientam (questionários lexicais para realização de atlas linguísticos, organizados por campos semânticos), registam variedades linguísticas em situação de alguma familiaridade entre inquiridores e informantes que quase sempre se envolvem emocionalmente, ao falarem sobre assuntos que conhecem bem e ao contarem, espontaneamente, as suas histórias pessoais. A caracterização social dos informantes é constante nas diferentes localidades do *corpus*, correspondendo ao perfil do informante tradicional dos trabalhos de dialectologia: idoso, pouco escolarizado ou analfabeto, trabalhador rural, natural da localidade e nela residente.

Mapa 1: Distribuição geográfica das localidades/micro-regiões do CORDIAL-SIN
(para legenda, ver Anexo)

[2] Não está encerrada nem pode ser aqui resumida a discussão das vantagens e dos inconvenientes dos diferentes tipos de metodologias empíricas que têm acompanhado o desenvolvimento da sintaxe dialectal nas últimas décadas. Para alguma informação de síntese, vejam-se Cornips & Poletto (2005), Barbiers (2008) e o capítulo relativo a metodologias de recolha de dados do *Dialect Syntax*

O *corpus* é disponibilizado através da página *internet* do CLUL (mais precisamente, em http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php), sob diferentes formatos: (i) transcrição ortográfica conservadora, com marcações diversas de fenómenos de oralidade e de variantes fonéticas e morfológicas de potencial relevância para a anotação sintáctica[3]; (ii) transcrição "normalizada", que serve de base à anotação do *corpus* e que é resultante da eliminação dos fenómenos marcados na transcrição conservadora; (iii) texto etiquetado por palavra (etiquetagem morfossintáctica). No futuro, será igualmente disponibilizada a anotação sintáctica do *corpus*, actualmente em fase de desenvolvimento.

A anotação deste *corpus* dialectal visa permitir o acesso rápido e sistemático a informação morfológica e sintáctica precisa e detalhada, sobre dados linguísticos pouco conhecidos e, consequentemente, pouco trabalhados. A constituição do *corpus* e a anotação deste tipo de dados, que pressupõe a investigação de aspectos linguísticos inexplorados nos estudos sobre o português, estão intrinsecamente associadas ao desenvolvimento dos estudos de sintaxe dialectal portuguesa[4].

A anotação morfossintáctica (por palavra), já acessível, constituiu um primeiro passo para a disponibilização deste tipo de informação, a reforçar e enriquecer com a anotação sintáctica. Para além do formato de disponibilização integral do *corpus*, por localidade, adoptado para o CORDIAL-SIN (permitindo pesquisas dirigidas pelo utilizador, em função dos seus interesses e dos recursos de pesquisa automática com os quais esteja familiarizado), está também em preparação o acesso à anotação por palavra deste *corpus* através de uma ferramenta de pesquisa em linha, concebida e implementada no âmbito do projecto EDISYN para pesquisas, articuladas ou isoladas, sobre recursos com informação morfossintáctica dialectal de diferentes domínios linguísticos (CORDIAL-SIN, ASIS, DynaSAND, NDC e EMK)[5,6].

Manual em http://www.dialectsyntax.org/index.php/manual-mainmenu-67/chapter-2-methodology. Sobre o papel dos dados naturalísticos de um *corpus* dialectal nos estudos de sintaxe, cf. Carrilho (no prelo).

[3] Sobre a natureza e as convenções deste tipo de transcrição no CORDIAL-SIN, cf. Magro (2007).

[4] Os principais estudos produzidos no âmbito do CORDIAL-SIN estão disponíveis em: http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_publicacoes.php.

[5] Mais informação em: http://www.dialectsyntax.org/index.php/edisyn-othermenu-51.

[6] O sistema de anotação morfossintáctica (por palavra) do CORDIAL-SIN tem por base o sistema definido para o *Tycho Brahe Parsed Corpus of Historical Portuguese* e aplicado pelo etiquetador automático desenvolvido para este *corpus* (cf. Finger 1998, 2000); este etiquetador automático é igualmente utilizado no processo de etiquetagem por palavra do CORDIAL-SIN. Sobre as etiquetas morfossintácticas, suas aplicações e exemplos do CORDIAL-SIN, cf. Magro e Morgado (2008).

2. A anotação sintáctica do *corpus*

O sistema adoptado para a anotação do CORDIAL-SIN é um sistema de anotação refinada, que permite codificar informação sintáctica relevante e que produz representações estruturais automatica e exaustivamente pesquisáveis, por meio de um motor de busca para *corpus* anotado – *CorpusSearch2* (Randall, 2005-2007). Este sistema é baseado no sistema de anotação concebido originalmente para os *corpora* históricos do inglês da Universidade da Pensilvânia (Kroch e Taylor, 2000; Kroch, Santorini e Delfs, 2004; Kroch, Santorini e Diertani, 2010), que tem de ser naturalmente adaptado e/ou expandido para responder às necessidades particulares de anotação de um corpus de dados orais dos dialectos do português.

Nesta secção, centrada em questões relativas à anotação sintáctica do *corpus*, apresentamos (i) os princípios gerais do sistema de anotação adoptado, (ii) a metodologia seguida no processo de anotação e (iii) as ferramentas utilizadas para a anotação e pesquisa dos dados. Referimos ainda, numa perspectiva linguística, aspectos relativos à operabilidade de um sistema como o dos *Penn corpora* na anotação de dados do português dialectal, mostrando quais são as adaptações que o sistema original tem de sofrer para se adequar a este novo objectivo.

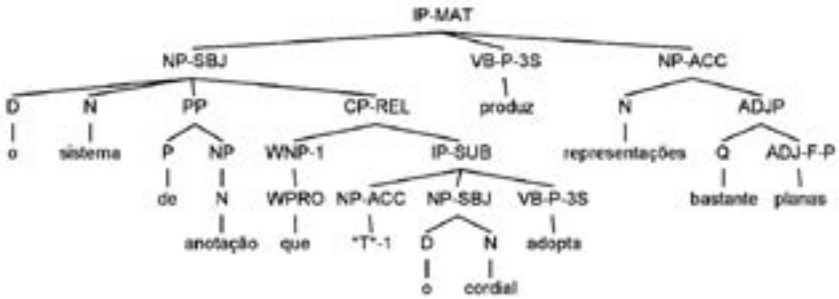
2.1. O sistema de anotação adoptado

O sistema de anotação sintáctica adoptado foi desenhado, no âmbito dos *corpora* históricos do inglês, para permitir a pesquisa automática de construções sintácticas consideradas relevantes. A atribuição de uma descrição estrutural correcta a cada frase do *corpus* não é, pois, um dos propósitos deste sistema.

Esta orientação leva a que as representações estruturais produzidas sejam bastante planas e, por vezes, linguisticamente descomprometidas, incluindo, por exemplo, nós de ramificação múltipla (e não necessariamente binária) e palavras que não projectam, como é o caso dos verbos, da negação e das partículas focalizadoras, entre outras. A não codificação de informação que é discutível, como, por exemplo, informação relativa a fronteiras de VP ou a distinções entre PPs complementos e adjuntos, ou a utilização de soluções "por defeito", como acontece com a localização de categorias vazias, são outros dos factores que contribuem para o aspecto teoricamente não motivado das representações.

A representação em (1) ilustra o tipo de estrutura produzida por este sistema.

(1)



Não obstante, este sistema de anotação permite realmente uma anotação refinada, que codifica fronteiras de constituinte, dependências sintagmáticas e oracionais, informação categorial (e.g. NP, PP, ADVP, IP), relações gramaticais (SBJ, ACC, DAT), funções discursivas (e.g. tópicos, marcadores pragmáticos), tipos de frase/oração (e.g. exclamativas, comparativas, interrogativas), alguns constituintes nulos e algumas relações transformacionais.

A anotação sintáctica é implementada sobre o texto morfologicamente etiquetado, sob a forma de parentetização etiquetada. Ao nível da palavra as etiquetas morfológicas mantêm-se; ao nível do sintagma ou da oração/frase, as etiquetas principais são etiquetas categoriais e as subetiquetas dão informação relativa a subcategoria, relação gramatical ou função discursiva.

A Tabela 1 apresenta o conjunto nuclear das etiquetas e subetiquetas sintácticas do sistema original.

Etiquetas Sintagmáticas	
NP	Sintagma Nominal
NP-SBJ	Sintagma Nominal (sujeito)
NP-ACC	Sintagma Nominal (OD)
NP-ADV	Sintagma Nominal (adverbial)
NP-VOC	Sintagma Nominal (vocativo)
NP-DAT	Sintagma Nominal (OI)
NP-GEN	Sintagma Nominal (dativo de posse)
PP	Sintagma Preposicional
PP-ACC	Sintagma Preposicional (complemento partitivo)
ADVP	Sintagma Adverbial
ADJP	Sintagma Adjectival
NUMP	Sintagma Numeral
INTJP	Sintagma Interjectivo
QP	Sintagma Quantificador
WXP	Sintagma-Wh (e.g. WNP, WPP)

Etiquetas Oracionais/Frásicas	
IP-MAT	IP Independente ou coordenado
IP-IND	IP Independente, não declarativo
IP-SUB	IP Subordinado (sob CP)
IP-ADV	IP Adverbial
IP-INF	Infinitiva
IP-GER	Gerundiva
IP-PPL	Participial
IP-SMC	Oração Pequena
CP-THT	Completiva Finita
CP-REL	Relativa
CP-FRL	Relativa Livre
CP-CLF	Clivada
CP-ADV	Adverbial
CP-DEG	Consecutiva
CP-CMP	Comparativa
CP-EXL	Exclamativa
CP-IMP	Imperativa
CP-QUE	Interrogativa

Tabela 1: Principais etiquetas sintagmáticas e oracionais/frásicas

Os exemplos (2) e (3) ilustram os dois níveis de anotação do CORDIAL-SIN (etiquetagem morfológica e anotação sintáctica, respectivamente) sobre uma mesma frase do *corpus*. A informação introduzida em cada versão do *corpus* está representada a negrito. Nas estruturas com parentetização etiquetada, como a de (3), o nível de indentação corresponde ao nível de encaixe estrutural.

(2)

e/**CONJ** andávamos/**VB-D-1P** com/P as/**D-F-P** redes/**N-P** @de/**P** @o/**D** badejo/**N**,/**,** que/**WPRO** são/**SR-P-3P** mais/**ADV-R** baixas/**ADJ-F-P** .../**.**

(3)

(IP-MAT (CONJ e)
(NP-SBJ *pro*)
(VB-D-1P andávamos)
(PP (P com)
(NP (D-F-P as)
(N-P redes)
(PP (P @de)
(NP (D @o)

(N badejo)
(, ,)
(**CP-REL (WNP-1** (WPRO que))
(**IP-SUB (NP-SBJ *T*-1**)
(SR-P-3P são)
(ADJP (ADV-R mais)
(ADJ-F-P baixas**)))))))**
(. ...)**)** [CORDIAL-SIN, VPA07]

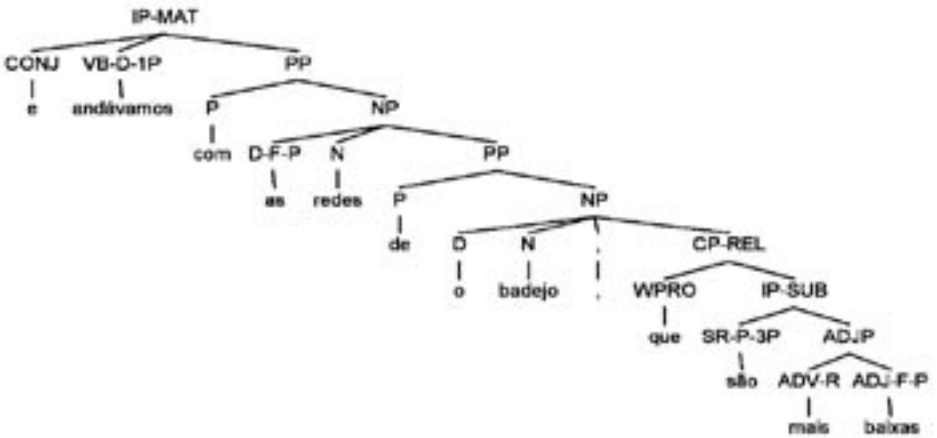
2.2. O processo de anotação dos dados

A anotação sintáctica do CORDIAL-SIN resulta de um processo em duas fases, desenvolvido com recurso a ferramentas dos *Penn Corpora.* Os textos etiquetados morfologicamente são, numa primeira fase, automaticamente segmentados por um analisador sintáctico de base estatística (analisador de Collins (1999) e Bikel (2004), especificamente modificado por Seth Kulick para a construção do *Penn Treebank*); numa segunda fase, o *output* do analisador automático é corrigido manualmente com o apoio do *CorpusDraw*, uma ferramenta de edição da anotação.

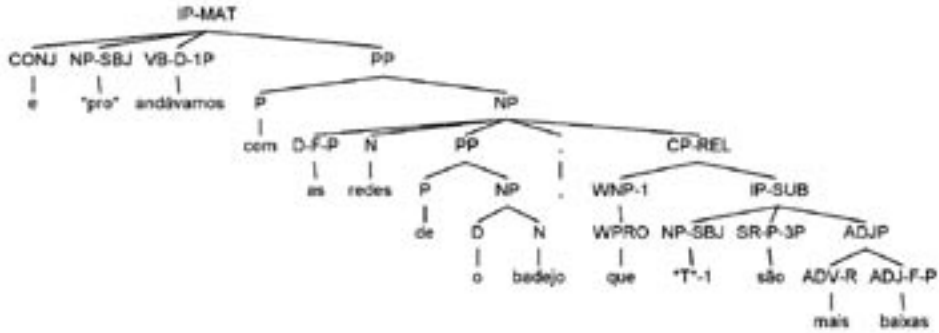
O *output* do analisador automático corresponde a um ficheiro regular de texto (ficheiro ASCII) contendo frases anotadas sob a forma de parentetização etiquetada (cf. exemplo (3)). Este ficheiro pode ser aberto através do *CorpusDraw*, caso em que a anotação sintáctica das frases é apresentada sob a forma de representação em árvore. Esta interface gráfica da ferramenta não só apresenta a anotação sintáctica de forma mais amigável, como permite ao anotador editar de forma rápida e controlada as representações apresentadas. As funções de edição da ferramenta permitem alterar etiquetas sintácticas, incluir informação subcategorial, corrigir o nível de encaixe estrutural, introduzir categorias vazias e coindexar constituintes movidos e categorias vazias correspondentes.

As estruturas apresentadas em (4) e (5) ilustram as duas etapas de anotação dos dados do CORDIAL-SIN: a anotação da frase em (4) foi atribuída pelo analisador automático; (5) representa a anotação da mesma frase após a edição manual da anotação. Note-se que, em (4), nenhuma das orações tem uma posição de sujeito anotada, o nível de encaixe da relativa está errado (a relativa está a modificar o nome *badejo* e não o nome *redes*, como deveria) e o sintagma-wh não está projectado nem co-indexado com o sujeito da relativa. Em (5) estas correcções foram introduzidas, por meio do *CorpusDraw*, e a frase ficou anotada de acordo com o sistema previsto.

(4)



(5)



2.3. A pesquisa de dados anotados

A versão com anotação sintáctica do CORDIAL-SIN poderá ser automaticamente pesquisada, por configuração estrutural, através do *CorpusSearch* (ver nota 7). O *CorpusSearch* é um motor de busca para *corpora* anotados que opera de forma amigável, utilizando uma linguagem de *query* básica e linguisticamente intuitiva e oferecendo opções de busca muito versáteis.

Um ficheiro de *query* do *CorpusSearch* (ilustrado em (6)) é constituído por um *node* (que indica o domínio da pesquisa) e um *query* (que indica o objecto da pesquisa). Uma expressão de *query* básica é composta por uma função de busca (correspondente a uma relação estrutural) e respectivos argumentos (constituintes a pesquisar). Os argumentos de uma função de busca podem ser coordenados disjuntivamente, podem incluir *wildcards* e podem ser negados. Duas ou mais funções de busca podem ser combinadas através dos operadores lógicos AND, OR e NOT, dando origem a *queries* mais complexos.

O *output* de uma busca feita com o *CorpusSearch* é um ficheiro de texto que reúne as frases que verificam as condições formuladas no ficheiro de *query*. O ficheiro de *output* de uma busca pode constituir o *input* da busca seguinte, o que permite restringir o âmbito das pesquisas efectuadas sem ter de escrever expressões de *query* complexas ou muito longas.

A Tabela 2 lista e descreve sumariamente as principais funções de busca do *CorpusSearch*.

Funções de Busca	**Descrição**
CCommands	nem x domina y, nem y domina x e o primeiro nó ramificado que domina x domina y
Dominates	x domina y
iDominates	x domina imediatamente y
iDomsFirst	y é o constituinte imediato de x mais à esquerda
iDomsLast	y é o constituinte imediato de x mais à direita
iDomsOnly	y é o único constituinte imediato de x
iDomsTotal	x tem um número determinado de constituintes imediatos
iDomsMod	x domina y, e o constituinte que intervem entre x e y é z
iDomsViaTrace	x domina imediatamente t e t está coindexado com um constituinte z
HasSister	x e y são imediatamente dominados pelo mesmo nó
Precedes	x antecede e não domina y
iPrecedes	x antecede imediatamente y e não domina y
SameIndex	x tem o mesmo índice de y
IsRoot	x é o nó raiz
Exists	x existe

Tabela 2: Principais funções de busca do *CorpusSearch*

Em (6) apresentamos o exemplo de um *query* básico para pesquisa de dados de redobro do clítico. Este *query* deverá ler-se da seguinte forma: sob qualquer tipo de sintagma nominal (NP*), procure casos em que um clítico (CL) tenha como nó irmão um sintagma preposicional (PP). Uma frase como a de (7) seria recuperável pelo *query* de (6).

(6)
node: NP*
query: (CL HasSister PP)

(7)
(IP-MAT (CONJ E)
    (NP-SBJ (PRO eu))

(VB-D-1S @fazia)
(NP-DAT (**CL @lhe**)
(**PP *ICH*-2)**)
(NP-ACC (N falta))
(**PP-2** (P @a)
(NP (D @o)
(PRO$ meu)
(N pai)))
(. .)) [CORDIAL-SIN, VPA15]

2.4. A adaptação do sistema existente para anotação de dados orais dos dialectos do português europeu

A anotação sintáctica de dados orais do português dialectal com recurso a um sistema de anotação refinada pré-existente, de aplicação já definida para *corpora* de inglês, implicou a adaptação e/ou expansão de alguns dos esquemas de representação originais, trabalho desenvolvido em colaboração com as equipas dos *Penn Corpora* e do *Tycho Brahe Parsed Corpus of Historical Portuguese*. As novas soluções de anotação cobrem, por um lado, casos de variação entre o português e o inglês ou entre variedades do português e, por outro, construções recorrentes na oralidade e praticamente ausentes dos registos escritos. As alterações ao sistema original são sempre minimizadas, procurando-se soluções compatíveis com as convenções estabelecidas e úteis do ponto de vista da pesquisa.

Os esquemas de anotação existentes são preservados sempre que possível, como acontece, por exemplo, com a anotação de uma completiva finita padrão (como (8)) e de uma completiva finita com duplicação do complementador (como em (9)):

(8)
(IP-MAT-PRN (NP-SBJ *exp*)
(VB-P-3S parece)
(**CP-THT** (**C** que)
(**IP-SUB** (NP-SBJ *pro*)
(SR-D-3S era)
(NP-PRD (N pintor)))))
[...] [CORDIAL, AAL04]
(9)
(IP-MAT (NP-SBJ *exp*)
(VB-P-3S Parece)
(**CP-THT** (**C** que)
(PP-1 (P @em)
(NP (D-F @a)

(NPR Suíça)))
(**CP-THT** (**C** que)
(**IP-SUB** (PP *ICH*-1)
(NP-SBJ *pro*)
(VB-P-3P dão)
(NP-ACC (Q-F muita)
(N importância))
(PP (P a)
(NP (D-F-P essas)
(N-P coisas))))))
(. .)) [CORDIAL, AAL04]

Noutros casos, a anotação do CORDIAL-SIN adapta esquemas existentes a novas possibilidades apresentadas pelos dados dialectais, como acontece com as orações relativas. Os exemplos (10) e (11) reproduzem o esquema de anotação previsto para relativas e relativas livres, respectivamente (CP-REL ou CP-FRL domina imediatamente um sintagma-wh – WNP, nos exemplos – e IP-SUB, que contém uma categoria vazia co-indexada com esse sintagma):

(10)
(IP-MAT [...]
(NP-SBJ (D-P os)
(N-P coelhos)
(**CP-REL (WNP-1** (WPRO que))
(**IP-SUB** (**NP-SBJ *T*-1**)
(VB-P-3P comem)
(NP-ACC (DEM isto)))))
(VB-P-3P morrem)
(. .)) [CORDIAL, AAL01]

(11)
(IP-MAT (ADVP (ADV Também))
(NP-SBJ *exp*)
(HV-P-3S há)
(NP-ACC (**CP-FRL** (**WNP-1** (WPRO quem))
(**IP-SUB** (**NP-SBJ *T*-1**)
(VB-SP-3S faça)
(NP-ACC (D-F essa)
(N coisa)))))
(. .)) [CORDIAL, AAL02]

A anotação em (12) adopta, para o CORDIAL-SIN, o esquema já previsto para casos de apagamento de preposição em relativas:

(12)
(IP-MAT (NP-SBJ-2 *exp*)
(SR-D-3S Era)
(NP-2 (D esse)
(N alqueive)
(**CP-REL** (**WPP**-4 (**P (CODE {em})**)
(WNP (WPRO que)))
(**IP-SUB** (**PP *T*-4**)
(NP-SBJ (D-F a)
(N gente))
(VB-P-3S lavra))))
[...] [CORDIAL, AAL14]

As relativas que incluem um elemento resumptivo recebem no CORDIAL-SIN uma anotação não originalmente prevista para relativas, recorrendo no entanto à combinação de parte do esquema de anotação das relativas com a marcação de um elemento -RSP (previsto no sistema original para anotação de elementos resumptivos de constituintes deslocados à esquerda):

(13)
(IP-MAT (CONJ e)
(VB-P-3S fica)
(NP-SBJ (D-F aquela)
(N coisa)
(**CP-REL** (**WNP** (WPRO que))
(**IP-SUB** (NP-SBJ-1 *exp*)
(ADV-NEG nunca)
(NP-SE-1 (CL se))
(**NP-OBL-RSP** (CL lhe))
(VB-P-3S mexe))))
(. .)) [CORDIAL, AAL15]

Noutros casos, como os das construções de marcação de tópico ou das construções de clivagem, a anotação do CORDIAL-SIN elabora e expande um esquema de anotação único inicialmente previsto, de forma a anotar de modo não-ambíguo as diferentes realizações que estas construções apresentam nos dados dialectais do português (cf. Carrilho & Magro, 2009).

Finalmente, o conjunto de etiquetas e subetiquetas do sistema inicial foi expandido pela inclusão de subetiquetas que marcam mais especificamente unidades sintácticas recorrentes em textos orais, com funções eminentemente discursivas (e que recebem uma anotação interna simplificada - por exemplo, sem marcação de categorias nulas). Nos exemplos seguintes apresenta-se a anotação de respostas (cf. subetiqueta -ANS(wer) nos exs. (14) e (15)), interrogativas-*tag* (cf. subetiqueta -TAG no ex. (16)), elementos de reforço de asserção (cf. subetiqueta –POL(arity) no ex. (17)) e diferentes marcadores pragmáticos (no sentido de Fraser (1996)) (cf. subetiquetas -ELAB(oration), -EDTS, para 'edit sign', e -PRG, para 'pragmatics' em geral, nos exs. (18) a (20)):

(14)
(INQ Passavam por sítios onde sabia que não havia guarda, não é?)
(**IP-ANS** (ADVP (ADV Pois))
(. .)) [CORDIAL, AAL66]

(15)
(INQ1 E trazia-as já feitas?)
(**IP-ANS** (VB-D-1S Trazia)
(. .)) [CORDIAL, PFT12]

(16)
(IP-MAT (NP-SBJ *pro*)
(VB-D-1S @Amostrei)
(NP-DAT (CL @lhe))
(NP-ACC *ICH*-252)
(ADVP (ADV ontem))
(DS -)
(**CP-QUE-TAG** (NEG não)
(VB-D-1S amostrei)
(. ?))
(DS -)
(**, ,**)
(NP-252 (D-F a)
(N lula))
(. .)) [CORDIAL, VPA36]

(17)
(IP-MAT (NP-SBJ *pro*)
(NEG não)
(SR-P-3S é)

```
      (PP(P @de)
        (NP (D @este)
          (N género)))
      (, ,))
(IP-POL (NEG não)
      (. .))  [CORDIAL, VPA07]

(18)
(IP-MAT-SPE (QT “)
        (NP-SBJ-1 (D-F esta)
              (N tapada))
        (, ,)
        (CONJ-EDTS ou)
        (NP-ELAB-1 (D este)
              (N curral))
        (, ,)
        (TR-P-3S tem)
        (NP-ACC (ADJ-R-P tantos)
              (N-P alqueires)
              (PP (P de)
                (NP (N trigo)
                  (CONJ ou)
                  (N centeio))))
        (QT “))
        (. .))  [CORDIAL,AAL07]
(19)
(IP-MAT (NP-SBJ *pro*)
      (ADVP-PRG (ADV bom))
      (, ,)
      (VB-P-3P fazem)
      (ADVP (ADV assim))
      (. .))  [CORDIAL, PST01]

(20)
(IP-MAT (NP-LFD (D-F A)
            (N pesca))
      (, ,)
      (CP-IMP-PRG (VB-SP-3S olhe))
      (, ,)
      (VB-P-3S @larga)
```

```
(NP-SE-14 CL @se))
(NP-SBJ-14 (D-F a)
        (N rede))
(PP (P por)
   (NP (D-F a)
      (N borda)))
(. .)) [CORDIAL, VPA05]
```

3. Conclusão

A versão do CORDIAL-SIN com anotação por palavra, actualmente disponível, integra já informação morfossintáctica de algum detalhe e possibilita uma pesquisa bastante apurada dos dados, quando explorada por ferramentas adequadas como, por exemplo, um gerador de concordâncias. A versão com anotação sintáctica representa, porém, um outro estádio de refinamento da anotação dos dados, permitindo, entre outras coisas, a codificação de material sintacticamente relevante ausente dos textos etiquetados, como os constituintes nulos. Com este tipo de anotação, o CORDIAL-SIN passará a ser automaticamente pesquisável não só por palavra ou por etiqueta morfossintáctica mas por configuração estrutural, tornando-se um recurso mais eficaz para a investigação em sintaxe dialectal.

É ainda de salientar que a própria tarefa de anotação dos dados faz identificar construções sintácticas dialectais que, de outra forma, passariam despercebidas. A anotação sintáctica pode, assim, orientar o trabalho futuro de inspecção do *corpus* e acelerar a divulgação de um vasto conjunto de fenómenos dialectais, inexistentes na variedade *standard* mas relevantes para o estudo da sintaxe do português em geral.

Referências

ALEAç – Atlas Linguístico e Etnográfico dos Açores (J. Saramago, coord.) [http://www.clul.ul.pt/sectores/variacao/projecto_aleac.php]

ALLP – Atlas Linguístico do Litoral Português (G. Vitorino, coord.) [http://www.clul.ul.pt/sectores/variacao/projecto_allp.php]

ALEPG – Atlas Linguístico-Etnográfico de Portugal e da Galiza (J. Saramago, coord.) [http://www.clul.ul.pt/sectores/variacao/projecto_alepg.php]

ASIS – Atlante Sintattico d'Italia (P. Benincà, coord.) [http://asis-cnr.unipd.it/]

BA – Segura, Maria Luísa (1987). *A Fronteira Dialectal do Barlavento do Algarve*. Dissertação de Doutoramento, Universidade de Lisboa.

Barbiers, Sjef (2008). Locus and limits of syntactic microvariation. *Lingua* 199, pp. 1607-1623.

Bikel, Daniel (2004). *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. PhD Dissertation, University of Pennsylvania.

Carrilho, Ernestina (no prelo). Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects). In G. Aurrekoetxea e J. L. Ormaetzea (eds.) *Tools for Linguistic Variation*. Bilbao: ASJU-ren gehigarriak, LIII, UPV-EHU.

Carrilho, Ernestina & Catarina Magro (2009). *CORDIAL-SIN – Syntactic Annotation System Manual*. [http://www.clul.ul.pt/english/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html]

Collins, Michael (1999). *Head-Driven Statistical Models for Natural Language Processing*. PhD Dissertation,. University of Pennylvania.

CORDIAL-SIN – The Syntax-oriented Corpus of Portuguese Dialects (A. M. Martins, coord.)[http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php]

Cornips, Leonie & Cecilia Poletto (2005). On Standardizing Syntactic Elicitation Techniques (part 1). *Lingua* 115, pp. 939-957.

DynaSAND – Dynamische Syntactische Atlas von de Nederlandse Dialecten (S. Barbiers, coord.) [http://www.meertens.knaw.nl/sand]

EDISYN – European Dialect Syntax Project (S. Barbiers, coord.) [http://www.dialectsyntax.org/index.php/project-description-edisyn-mainmenu-50]

EMK – Estonian Dialect Corpus (L. Lindström, coord.) [http://www.murre.ut.ee/triip/home/]

Finger, Marcelo (1998). Tagging a morphologically rich language. In *Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98)*. Brno: Masaryk University Press, pp. 39-44.

Finger, Marcelo (2000). Técnicas de Otimização da precisão empregadas no etiquetador Tycho Brahe. In *Actas do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*. São Paulo: ICMC/USP, pp. 141-154.

Fraser, Bruce (1996). Pragmatic Markers. *Pragmatics* 6 (2), pp. 167-190.

Kroch, Anthony & Anne Taylor (2000). *Penn-Helsinki Parsed Corpus of Middle English*, second edition. [http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html]

Kroch, Anthony, Beatrice Santorini & Lauren Delfs (2004). *Penn-Helsinki Parsed Corpus of Early Modern English*. [http://www.ling.upenn.edu/emodeng]

Kroch, Anthony, Beatrice Santorini & Ariel Diertani (2010). *Penn Parsed Corpus of Modern British English*. [http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html]

Magro, Catarina (org.) (2007). *CORDIAL-SIN. Normas de transcrição.* [http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php]

Magro, Catarina & Cristina Morgado (orgs.) (2008). *CORDIAL-SIN. Manual de Anotação Morfossintáctica*. [http://www.clul.ul.pt/sectores/variacao/cordialsin/manual_anotacao_morfologica.pdf]

NDC – Nordic Dialect Corpus (J. Johannessen, J. Priestley, K. Hagen, T. Åfarli & Ø. Vangsnes, coords.) [http://www.tekstlab.uio.no/nota/scandiasyn/index.html]

Randall, Beth (2005-2007). *CorpusSearch 2*. [http://corpussearch.sourceforge.net]

Tycho Brahe Parsed Corpus of Historical Portuguese (C. Galves, coord.). [http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html]

Anexo: Lista de localidades/micro-regiões do CORDIAL-SIN

01	VPA	Vila Praia de Âncora (Viana do Castelo)	**22**	EXB	Enxara do Bispo (Lisboa)
02	CTL	Castro Laboreiro (Viana do Castelo)	**23**	TRC	Fontinhas (Angra do Heroísmo)
03	PFT	Perafita (Vila Real)	**24**	MTM	Moita do Martinho (Leiria)
04	AAL	Castelo de Vide, Porto da Espada, S. Salvador de Aramenha, Sapeira, Alpalhão, Nisa (Portalegre)	**25**	LAR	Larinho (Bragança)
05	PAL	Porches, Alte (Faro)	**26**	LUZ	Luzianes (Beja)
06	CLC	Câmara de Lobos, Caniçal (Funchal)	**27**	FIS	Fiscal (Braga)
07	PST	Camacha, Tanque (Funchal)	**28**	GIA	Gião (Porto)
08	MST	Monsanto (Castelo Branco)	**29**	STJ	Santa Justa (Santarém)
09	FLF	Fajãzinha (Horta)	**30**	UNS	Unhais da Serra (Castelo Branco)
10	MIG	Ponta Garça (Ponta Delgada)	**31**	VPC	Vila Pouca do Campo (Coimbra)
11	OUT	Outeiro (Bragança)	**32**	GRJ	Granjal (Viseu)
12	CVB	Cabeço de Vide (Portalegre)	**33**	CRV	Corvo (Horta)
13	MIN	Arcos de Valdevez, Bade, São Lourenço da Montaria (Viana do Castelo)	**34**	GRC	Graciosa (Angra do Heroísmo)
14	FIG	Figueiró da Serra (Guarda)	**35**	MLD	Melides (Setúbal)
15	ALV	Alvor (Faro)	**36**	STA	Santo André (Vila Real)
16	SRP	Serpa (Beja)	**37**	MTV	Montalvo (Santarém)
17	LVR	Lavre (Évora)	**38**	CLH	Calheta (Angra do Heroísmo)
18	ALC	Alcochete (Setúbal)	**39**	CPT	Carrapatelo (Évora)
19	COV	Covo (Aveiro)	**40**	ALJ	Aljustrel (Beja)
20	PIC	Bandeiras, Cais do Pico (Horta)	**41**	STE	Santo Espírito (Ponta Delgada)
21	PVC	Porto de Vacas (Coimbra)	**42**	CDR	Cedros (Horta)