European Dialect Syntax III
Venice • 18-21 September

The CORDIAL Parsing

Ernestina Carrilho
Catarina Magro
(University of Lisbon)

CLUL Centro de Linguística da Universidade de Lisboa

Overview of presentation

- short description of CORDIAL-SIN (corpus properties and current state)
- detailed account of CORDIAL-SIN syntactic annotation
  - the annotation system – goals, principles and guidelines
  - the annotation process – tools and procedures
  - the adjustments on the system required by a Portuguese corpus of dialectal speech

CLUL

CORDIAL-SIN
Corpus Dialectal para o Estudo da Sintaxe

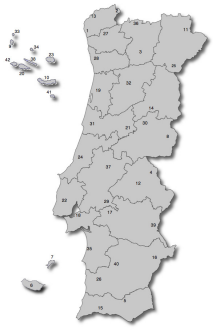
- Syntax-oriented Corpus of Portuguese Dialects

- a corpus of spoken dialectal European Portuguese
- a geographically representative body of excerpts of spontaneous and semi-directed speech
- excerpts selected from the oral interviews gathered by the Research Group on Linguistic Variation at CLUL within the scope of several linguistic atlases
  - Informants: - aged
    - low instruction
    - rural
    - born and raised in place of interview

CLUL

CORDIAL-SIN corpus

- cc. 600 000 words
- 42 locations



available online:

- ✓ *verbatim* orthographic transcripts
- ✓ 'normalized' orthographic transcripts
- ✓ morphologically tagged texts

http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php

CLUL

The CORDIAL-SIN project

- Building up the tagged corpus
  [from 1999 to 2007]
  national funding by *Fundação para a Ciência e Tecnologia* (FCT)
  - CORDIAL-SIN (PRAXIS XXI/P/PLP/13046/1998)
  - CORDIAL-SIN2 (POSI/1999/PLP/33275)
  - Dialect Syntax (POCTI/LIN/46980/2002)
- Corpus syntactic annotation
  [since January 2008]
  - DUPLEX (PTDC/LIN/71559/2006)

CLUL

CORDIAL-SIN Syntactic annotation

- Based on the processes and tools used by the Penn Parsed Corpora of Historical English (http://www.ling.upenn.edu/hist-corpora)
  - the *Penn-Helsinki Parsed Corpus of Middle English*, second edition [PPCME2] (Kroch & Taylor 2000)
  - the *Penn-Helsinki Parsed Corpus of Early Modern English* [PPCEME] (Kroch, Santorini & Delfs 2004)
  - the *Penn Parsed Corpus of Modern British English* [PPCMBE] (Kroch & Santorini, under construction)

CLUL

CORDIAL-SIN Syntactic annotation
Corpora network

- Other corpora using the Penn Corpora annotation scheme
  - *The Tycho Brahe Corpus* (a parsed corpus of historical Portuguese)
    (http://www.tycho.iel.unicamp.br/~tycho/)
    Charlotte Galves (University of Campinas, Brazil)
  - *The Canadian Parsed Corpus of Historical French*
    France Martineau (University of Ottawa, Canada)

CLUL

CORDIAL-SIN Syntactic annotation
The Penn Corpora annotation system
Goals and Principles

- to facilitate automated searches for various constructions of interest
- (!) not to associate every sentence with a correct structural description

Dealing with uncertainty and ambiguity
Avoid controversial decisions

- Omiting undecidable information
  - VP boundaries
  - subtle distinctions (adjectival vs. verbal passives, argument vs. adjunct PPs)
- Using default (and, sometimes, linguistically unmotivated) rules
  - location of wh-traces
  - PP attachment ("when in doubt, attach high")

CLUL

CORDIAL-SIN Syntactic annotation
The Penn Corpora annotation system
Goals and Principles

- passive *vs* impersonal *se* – a default rule option
  - unambiguous *se*

passive *se*

```
(IP-MAT   (VB-P-3P @Põem)
          (NP-SE-1  (CL @se))
          (NP-SBJ-1  (D-UM-P uns)
                     (N-P panos))
          (. .))

          [SRP24]
```

impersonal *se*

```
(IP-MAT   (CONJ mas)
          (NP-SBJ-1 *exp*)
          (FP já)
          (NP-SE-1 (CL se))
          (VB-D-3S vivia)
          (ADVP (ADV-R melhor)))

          [VPA15]
```

CLUL

CORDIAL-SIN Syntactic annotation
The Penn Corpora annotation system
Goals and Principles

- passive *vs* impersonal *se* – a default rule option
  - ambiguous *se*

passive *se*

```
(IP-MAT  (PP (P @A)
             (NP   (D-F-P @as)
                   (N-P vezes)))
         (VB-P-3S @apanha)
         (NP-SE-1 (CL @se))
         (NP-SBJ-1 (D-UM-F uma)
                   (N quantidade)
                   (PP (P @de)
                       (NP  (DEM @aquilo))))
         (. .))  [VPA10]
```

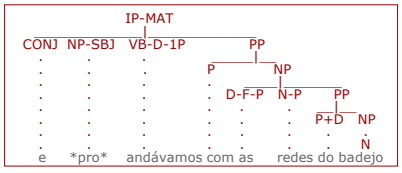
impersonal *se*

```
(IP-MAT  (NP-SBJ-1 *exp*)
         (PP (P @A)
             (NP   (D-F-P @as)
                   (N-P vezes)))
         (VB-P-3S @apanha)
         (NP-SE-1 (CL @se))
         (NP-SBJ (D-UM-F uma)
                 (N quantidade)
                 (PP (P @de)
                     (NP  (DEM @aquilo))))
         (. .))  [VPA10]
```

CLUL

CORDIAL-SIN Syntactic annotation
The Penn Corpora annotation system
Goals and Principles

- Quite 'flat' structures
  - multiple branching nodes
  - some word-level nodes (VB, NEG, FP, a.o.)



CLUL

CORDIAL-SIN Syntactic annotation
The Penn Corpora annotation system
General Guidelines

Although…

- Rich annotation system – marking up of:
  - constituent boundaries
  - phrase and clause dependencies
  - categorial information (e.g. NP, PP, ADVP)
  - grammatical functions (e.g. SBJ, ACC)
  - sentence and clause type (e.g. EXL, QUE)
  - some null constituents
  - some transformational relations
  - other relevant information (LFD, pragmatic markers)

CLUL

CORDIAL-SIN Syntactic annotation
Example

- syntactic annotation is represented as labeled bracketing
  (depth of indenting corresponds to depth of structural embedding)

- from morphologically tagged texts:

  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 …/. [VPA07]

- to syntactically annotated texts…

CORDIAL-SIN Syntactic annotation
Example

```
(IP-MAT (CONJ e)                                                    and
        (NP-SBJ *pro*)                                              –
        (VB-D-1P andávamos)                                         were1PL
        (PP (P com)                                                 with
            (NP (D-F-P as)                                          the
                (N-P redes)                                         fishing net
                (PP (P @de)                                         of
                    (NP (D @o)                                      the
                        (N badejo)                                  whiting
                    (, ,)                                           ,
                    (CP-REL (WNP-1 (WPRO que))                      that
                            (IP-SUB (NP-SBJ *T*-1)                  –
                                    (SR-P-3P são)                   are3PL
                                    (ADJP (ADV-R mais)              more
                                          (ADJ-F-P baixas))))))))   deep
        (. ...)) [VPA07]
```

CORDIAL-SIN Syntactic annotation
Labels and extended labels
Phrase labels

Label	Category (& Function)
NP	Noun Phrase
NP-SBJ	Noun Phrase (subject)
NP-ACC	Noun Phrase (DO, nominal predicate)
NP-ADV	Noun Phrase (adverbial)
NP-VOC	Noun Phrase (vocative)
NP-DAT	Noun Phrase (dative)
NP-GEN	Noun Phrase (dative of possession)
PP	Prepositional Phrase
PP-ACC	Prepositional Phrase (partitive object)

CORDIAL-SIN Syntactic annotation
Labels and extended labels
Phrase labels

Label	Category (& Function)
ADVP	Adverbial Phrase
ADJP	Adjective Phrase
NUMP	Numeral Phrase
INTJP	Interjection Phrase
QP	Quantifier Phrase
WXP	Wh-Phrase (e.g. WNP, WPP)

CORDIAL-SIN Syntactic annotation
Labels and extended labels
Phrase labels

Label	Category (& Function)
IP-MAT	Independent or conjoined declarative IP
IP-IND	Independent, non-declarative IP
IP-SUB	Subordinate IP (under CP)
IP-ADV	Adverbial IP
IP-INF	Infinitival clause
IP-GER	Gerund clause
IP-PPL	Participial clause
IP-SMC	Small clause
…	…

CORDIAL-SIN Syntactic annotation
Labels and extended labels
Phrase labels

Label	Category (& Function)
CP-THT	*That* clause
CP-REL	Relative
CP-FRL	Free Relative
CP-CLF	Cleft
CP-ADV	Adverbial clause
CP-DEG	Degree clause
CP-CMP	Comparative clause
CP-EXL	Exclamative
CP-IMP	Imperative
CP-QUE	Question
…	…

CORDIAL-SIN Syntactic annotation
The syntactic annotation **process**

➢ automatic parsing of the POS tagged corpus
version of the Collins/Bikel statistical parser
(Collins 1999, Bikel 2004), modified for treebank construction by Seth Kulick

➢ hand correction of the parser output
CorpusDraw – an editing annotation tool
component of CorpusSearch2 (Randall 2005-2007 http://corpussearch.sourceforge.net)

➢ make available a parsed version of the corpus that allows to retrieve syntactic configurations of interest
CorpusSearch – a query language for parsed corpora
component of CorpusSearch2 (Randall 2005-2007 http://corpussearch.sourceforge.net)

CLUL

CORDIAL-SIN Syntactic annotation
The syntactic annotation **tools**

➢ CorpusSearch2 (Randall 2005-2007) – a java program that

- supports research in corpus linguistics
- is useful both for the construction of syntactically parsed corpora and for searching them
- runs under any Java-supported operating system (Linux, Mac, Unix, Windows) – requires Java 2, version 1.5 or later
- expects labelled bracketing (Penn Treebank style)
- is freely available from http://corpussearch.sourceforge.net/
- has two main components
  - CorpusDraw
  - CorpusSearch

CLUL

CORDIAL-SIN Syntactic annotation
The syntactic annotation **tools**

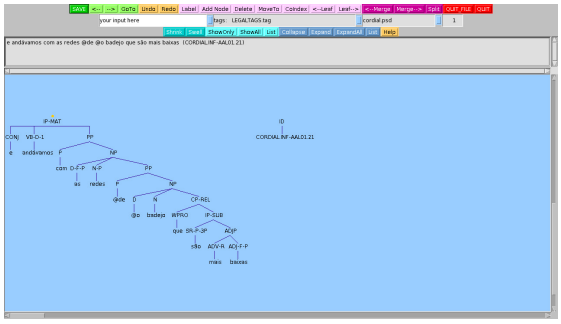
➢ CorpusDraw
Human editing of the parser output

Editing operations include:

- changing syntactic tags
- breaking up run-on sentences or consolidating fragments
- adding subcategory information
  - NP → NP-ADV
  - CP → CP-CMP
- changing attachment level
- adding empty categories
- adding indices to
  - an antecedent and its trace
  - a constituent in a noncanonical position and the position in which it is interpreted

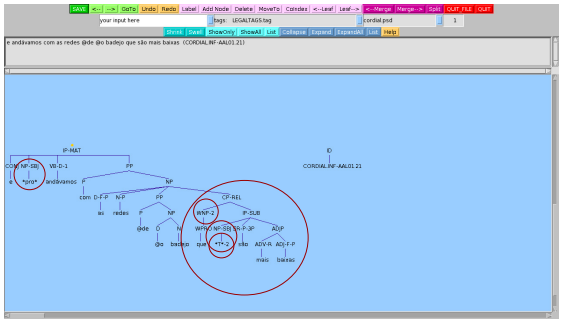
CLUL

CorpusDraw – parser output



CLUL

CorpusDraw – post-editing



CLUL

CORDIAL-SIN Syntactic annotation
The syntactic annotation **tools**

➢ CorpusSearch
a dedicated search engine for parsed corpora

Key features:

- basic search functions are linguistically intuitive
  - (immediately) precedes
  - (immediately) dominates
  - exists
  - hassister
  - isroot
  - logical operators: AND OR NOT
- end user can custom-define further linguistically relevant search expressions
- searches can disregard material as necessary
- the output of CorpusSearch is itself searchable

CLUL

CORDIAL-SIN Syntactic annotation
The syntactic annotation **tools**
CorpusSearch – query example

➢ A simple sample query

node: IP*
query: (IP* iDoms NEG)

- asterisk is a wildcard (IP* matches IP-MAT, IP-SUB, IP-INF, etc.)
- CorpusSearch searches the corpus for constituents with the label(s) specified in node
- whenever it finds such a constituent, it checks whether the material in the constituent matches the condition(s) in query
- matching tokens are recorded in an output file

CLUL

CORDIAL-SIN Syntactic annotation
At present

➢ training the Penn parser with European Portuguese data
(from July to December)
- parse a sample file of 10,000 words
- train the parser with the hand-corrected output
- repeat the process with new 10,000 words files up to a maximum of 50,000 words
- run the trained parser on the remaining 550,000 words

➢ defining the annotation system for European Portuguese
(in collaboration with Tycho Brahe and Penn Corpora teams)

➢ developing the annotator's manual
(http://www.clul.ul.pt/english/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html)

CLUL

CORDIAL-SIN Syntactic annotation
Prospects

➢ How long does it take to produce a parsed corpus of 600.000 words?

- mean editing speed: 500 words/hour
(in language well-known to annotator)
- 1 annotator
- annotators can work approx. 4 hours/day or 20 hours/week

600,000 words / 500 words/hour	=	1,200 hours
1,200 hours / 20 hours/week	=	60 weeks

CLUL

CORDIAL-SIN Syntactic annotation
Adapting the Penn system

- Annotation schemes
  - preserved wherever possible
  - adapted
  - expanded
  - (partly) omitted in specific domains
- Label set
  - preserved
  - new extended labels added

to parse
spoken
Portuguese
dialectal data

CLUL

Annotation scheme preserved
that clause complements

```
(IP-MAT  (CONJ e)                                          and
         (NP-SBJ (PRO ele))                                he
         (VB-D-3S disse)                                   said3SG
         (, ,)
         (ADV também)                                      also
         (, ,)
         (CP-THT  (C que)                                  that
                  (IP-SUB  (NP-SBJ *pro*)                  used3PL
                           (VB-D-3P usavam)
                           (NP-ACC *)))
         (. ...))   [VPA17]
```

CLUL

CP recursion

```
I        (IP-MAT  (NP-SBJ (PRO Eu))
know1SG           (VB-P-1S sei)
that              (CP-THT  (C que)
that                       (CP-THT  (DEM aquilo))
that                                (C que)
                                    (IP-SUB
not                                         (NEG não)
is                                          (SR-P-3S é)
for                                         (PP  (P por)
evil                                                (ADVP (ADV mal))))))
                  […]      [VPA15]
```

CLUL

Annotation scheme adapted
clitic climbing

```
(IP-MAT   (NP-SBJ  *pro*)
          (ADV também)                    also
          (NP-25 (CL o))                  it
          (VB-P-1S vou)                   go1SG
          (IP-INF    (NP-ACC *ICH*-25)
                     (VB levar))          take
          (. ...))   [VPA17]
```

CLUL

Annotation scheme expanded
relative clauses

preserved

```
a       ([...]  (NP-ACC   (D-UM-F uma)
woman                     (N senhora)
who                       (CP-REL          (WPRO que)
                                IP-SUB
had3SG                                     (TR-D-3S tinha)
seven                                      (NP-ACC  (NUM sete)
children                                             (N-P filhas)) ))
          (. .))          [PFT25]
```

CLUL

free relative

preserved

```
and   (IP-MAT   (CONJ e)
                (NP-SBJ *exp*)
has3SG          (HV-P-3S há)
who             (NP-ACC   (CP-FRL   (WNP-1 (WPRO quem))
                                    (IP-SUB  (NP-SBJ *T*-1)
                                             (VB-SP-3S largue)
throw                                        (NP-ACC   (D-F a)
the                                                    (N rede))
fishing net                                  (PP (P por)
from                                             (NP (D-F a)
the                                                  (N popa))))))
stern           (DS -)
                [...]          [VPA05]
```

'there are people who throw the fishing net from the stern'

CLUL

'chopping' relative

preserved

```
and   (IP-MAT   (CONJ E)
after           (ADVP (ADV depois))
there           (ADVP (ADV lá))
went3SG         (VB-D-3S foi)
the             (NP-SBJ   (D o)
boat                      (N barco)
{in}                      (WPP-167 (P  (CODE {em}))
which                                  (WNP (WPRO que)))
                          (IP-SUB   (PP *T*-167)
I                                   (NP-SBJ (PRO eu))
was                                 (VB-D-1S andava)) )
          (, ,))          [VPA28]
```

CLUL

resumptive relative

```
                                                  +
(NP  (D-F Essa)                                   that
     (ADJ-G tal)                                  such
     (N feiticeira)                               (fish name)
          (WNP (WPRO que))                        that
          (IP-SUB   (NP-SBJ *pro*)
                    (NP-DAT-RSP-28 (CL lhe))      it
                    (VB-P-3P chamam)              call3SG
                    (IP-SMC     (PP-SBJ *ICH*-28)
                                (NP-ACC (NP(D-F a)  the
                                             (N pata-roxa)
     [...]) [VPA11]
```

CLUL

Annotation scheme expanded
topic constructions

- Penn system: left dislocated noun-phrases (NP-LFD)

```
(IP-MAT   (NP-LFD     (NPR Mary))
          (NP-SBJ-RSP   (PRO she))
          (ADVP-TMP (ADV always))
          (VBP  wins)
          (. .))
```

CLUL

topic constructions in EP

(1) Mas esse peixe, já uma pessoa às vezes não o$_{ACC}$ conhece.
but that fish already a people sometimes not it know

(2) Nós tocava-nos$_{DAT}$ para aí uns quinhentos mil réis, ou setecentos ou oitocentos.
we ± belonged to.us about some five hundred thousand réis or seven hundred or eigth hundred

(3) E o lavagante já não há -, também.
and the lobster already not have$_{3SG}$ also 'there is no lobster either'

(4) A pesca, olhe, larga-se a rede por a borda.
the fishing look$_{SUBJ.3SG}$ throw SE the net from the stern

topic constructions in EP
clitic left dislocation

but	(IP-MAT	(CONJ Mas)	*preserved*
that		(**NP-LFD** (D esse)	
fish		(N peixe))	
		(, ,)	
		(FP já)	
		(NP-SBJ (D-UM-F uma)	
a		(N pessoa))	
person		(PP (P @a)	
		(NP (D-P @as)	
sometimes		(N-P vezes)))	
not		(NEG não)	
it		(NP-ACC-RSP (CL o))	
knows		(VB-P-3S conhece)	
		(. .)) [VPA30]	

topic constructions in EP
left dislocation

we	(IP-MAT	(NP-SBJ *exp*)	*preserved*
		(**NP-LFD** (PRO Nós))	
		(, ,)	
belonged		(VB-D-3S @tocava)	
us$_{DAT}$		(**NP-DAT-RSP** (CL @nos))	
		(NP-ACC (PP(P para)	
about		(ADVP (ADV aí)))	
some		(D-UM-P uns)	
five hundred		(NUMP (NUM quinhentos)	
thousands		(NUM mil))	
"réis"		(N-P réis))	
(old money)		[…]	
		(. .)) [VPA03]	

topic constructions in EP
"topicalization"

(IP-MAT	(CONJ E)		*and*
	(NP-SBJ *exp*)		*the*
	(**NP-8** (D o)		*lobster*
	(N lavagante))		
	(FP já)		
	(NEG não)		*not*
	(HV-P-3S há)		*has*
	(**NP-ACC *ICH*-8**)		*(existential)*
	(, ,)		
	(ADV também)		
	(. .)) [VPA01]		*also*

topic constructions in EP
hanging topic

(IP-MAT	(NP-SBJ-14 *exp*)	
	(**NP-LFD**(D-F A)	*the*
	(N pesca))	*fishing*
	(, ,)	
	(CP-IMP-PRG (VB-SP-3S olhe))	*look*
	(, ,)	
	(VB-P-3S @larga)	*throw*
	(NP-SE-14 (CL @se)	*SE*
	(NP-ACC (D-F a)	*the*
	(N rede))	*fishing net*
	(PP (P por)	*from*
	(NP (D-F a)	*the*
	(N borda)))	*stern*
	(. .)) [VPA05]	

Annotation scheme (partly) omitted
special cases

- No null categories
- No extended labels

inside:
- IP-ANS
- IP-POL
- CP-QUE-TAG
- IP / CP-PRG

In connection with pragmatics

Adapting the label set
New extended labels

- IP – ANS — answers to YN and *wh*-questions
- IP – POL — reinforcement of assertion
- CP-QUE – TAG — question-tag
- XP / IP / CP - PRG — pragmatic (generic)

In connection with pragmatics

CLUL

IP-ANS

(CODE INQ1 E trazia-as já feitas?) *I: And did you bring them already done*
(CODE INF) *Inf:*
(**IP-ANS** (VB-D-1S Trazia) *brought*
(. .)) [PFT12]

(CODE INQ2 E tinha umas coisas para respirar?) *I: And did it have any thing to breathe through?*
(CODE INF)
(**IP-ANS** (ADVP (ADV-NEG Não_senhora)) *Inf: no [<no madam]*)
(. .)) [PFT40]

CLUL

IP-POL

(IP-MAT (CONJ Porque)	*because*
(NP-SBJ *pro*)	
(SR-P-3S é)	*is*
(ADJP (ADJ branco))	*white*
(. .))	
(**IP-POL** (SR-P-3S É)	*is*
(. .)) [VPA24]	

CLUL

IP-POL

(IP-MAT (NP-SBJ *pro*)	
(NEG não)	*not*
(SR-P-3S é)	*is*
(PP (P @de)	*of*
(NP (D @este)	*this*
(N género)))	*kind*
(, ,))	
(**IP-POL** (NEG não)	*not*
(. .)) [VPA07]	

CLUL

CP-QUE-TAG

(IP-MAT (NP-SBJ *pro*)	
(TR-P-3P têm)	*have$_{3PL}$*
(NP-ACC (D-UM um)	*a*
(N bocadinho)	*little*
(PP (P de)	*of*
(NP (N ferrugem))))	*rust*
(, ,)	
(**CP-QUE-TAG** (NEG não)	*not*
(TR-P-3P têm))	*have$_{3PL}$*
(. ?)) [VPA36]	

CLUL

XP / CP / IP(-…) - PRG

(IP-MAT (NP-SBJ *pro*)	
(**ADVP-PRG** (ADV bem))	***well***
(, ,)	
(VB-P-3P fazem)	*do$_{3PL}$*
(ADVP (ADV assim))	*in.this.way*
(. .)) [PST01]	

CLUL

XP / IP / CP(-…) - PRG

(IP-MAT (ADVP(ADV antigamente))	*in the past*
(NP-SBJ-2 *exp*)	
(VB-D-3S @secava)	*dried*$_{3SG}$
(NP-SE-2 (CL @se))	*SE*
(NP-ACC (D o)	*the*
(N carapau)	*mackerel*
(DS -)	
(IP-PRG (VB-P-3S quer)	*wants*
(VB dizer))	*to say*
(NP-PRN (D o)	*the*
(N sorelo)))	"sorelo"
[…] [VPA11]	

XP / IP / CP(-…) - PRG

(IP-MAT (**CP-IMP-PRG** (VB-SP-3S Olhe))	***look***
(, ,)	
(NP-SBJ (D-F-P as)	*the*
(N-P batatas))	*potatoes*
(VB-D-3P vinham)	*grew*$_{3PL}$
(, ,)) [VPA06]	

XP / IP / CP(-…) - PRG

(IP-MAT (CONJ porque)	*because*
(NP-SBJ (D aquele)	*that*
(N senhor))	*man*
(SR-P-3S é)	*is*
(PP (P @de)	*of*
(NP (D-F @a)	*the*
(N religião)	*religion*
(PP (P @de)	*of*
(NP (D-F @a)	*the*
(N verdade)))))	*truth*
(, ,)	
(**CP-QUE-PRG** (VB-P-3S sabe))	***know***$_{3SG}$
(. ?)) [VPA15]	*('you know')*