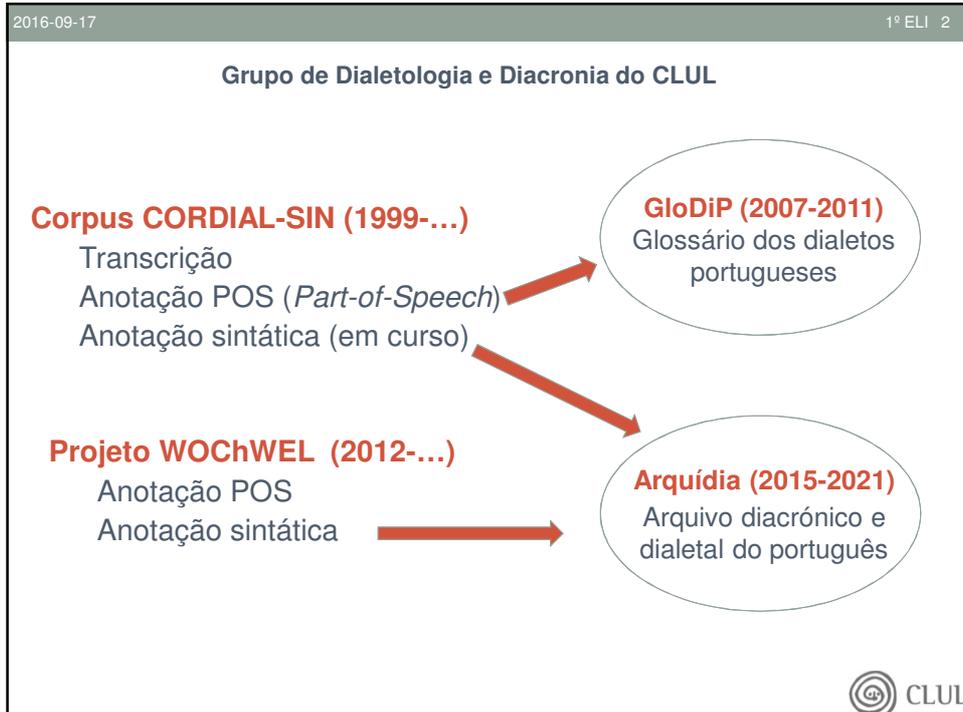


ARQUÍDIA

Arquivo diacrónico e dialetal do português

Sandra Pereira
sandra.pereira@campus.ul.pt

1º ELI – Encontro de Línguas Ibéricas
16-17 setembro, UBI - Covilhã



2016-09-17 1º ELI 3

Grupo de Dialetologia e Diacronia do CLUL

Corpus CORDIAL-SIN (1999-...)

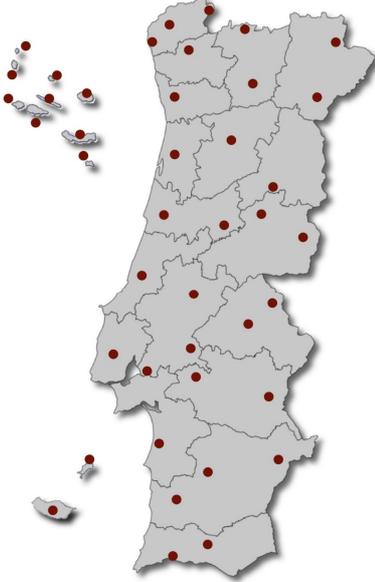
- Transcrição**
- Anotação POS (*Part-of-Speech*)**
- Anotação sintática (em curso)**

 CLUL

2016-09-17 1º ELI 4

CORDIAL-SIN: corpus dialetal para o estudo da sintaxe

- excertos de fala espontânea de inquiridos dialetais no âmbito de atlas linguísticos do CLUL (ALEPG, ALEAç,...)
- informantes:
 - não jovens
 - rurais
 - pouca escolaridade
 - nascidos e criados no lugar
- 600.000 palavras
- 70 horas
- 42 localidades
- transcrição conservadora
- transcrição normalizada
- anotação POS (*part-of-speech*)
- anotação sintática (em curso)



<http://www.clul.ul.pt/pt/recursos/212-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects>  CLUL

2016-09-17

1º ELI 5

CORDIAL-SIN

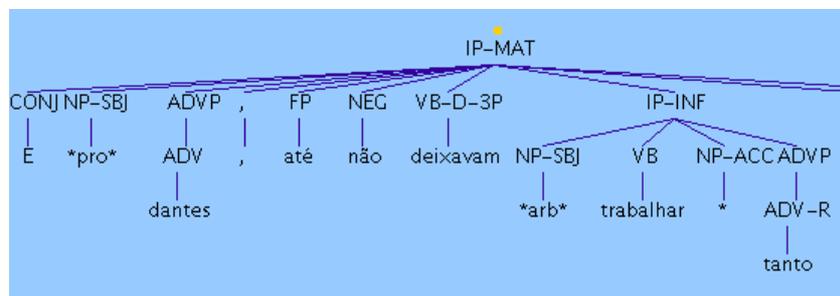
1. Estendem [AB|umas] uns sacos por baixo – agora é de plástico –, estendem uns plásticos por baixo, e depois vão lá acima, {PH|ku'mɛsĩ =começam} a baterem para baixo. (ALC17)
2. Estendem (...) uns sacos por baixo – agora é de plástico -, estendem uns plásticos por baixo, e depois vão lá acima, começam a baterem para baixo.
3. Estendem/VB-P-3P (...) uns/D-UM-P sacos/N-P por/P baixo/N - /DS agora/ADV é/SR-P-3S de/P plástico/N -/DS ,/, estendem/VB-P-3P uns/D-UM-P plásticos/N-P por/P baixo/N ,/, e/CONJ depois/ADV vão/VB-P-3P lá/ADV acima/ADV ,/, começam/VB-P-3P a/P baterem/VB-F-3P para/P baixo/N ./.



2016-09-17

1º ELI 6

CORDIAL-SIN



sistemas de anotação:
Tycho Brahe corpus; Penn corpora



2016-09-17 1º ELI 7

Grupo de Dialetologia e Diacronia do CLUL

Corpus CORDIAL-SIN (1999-...)

- Transcrição
- Anotação POS (*Part-of-Speech*)
- Anotação sintática (em curso)



GloDiP (2007-2011)
Glossário dos dialetos portugueses

 CLUL

2016-09-17 1º ELI 8

GloDiP: Glossário dos dialetos portugueses com informação sintática (Pereira 2012)

- Estabelecer o modelo de um glossário dos dialetos portugueses (informação lexical + informação sintática)
- Construir uma base de dados de verbos dos dialetos portugueses (= modelo para uma base de dados do léxico dialetal)
- Ampliar o conhecimento do léxico e da sintaxe dos dialetos portugueses, contribuindo para o crescimento da lexicografia portuguesa (lexicografia monolíngue baseada em *corpora*)
- Explorar as potencialidades de um *corpus* dialetal
- Contribuir para a interdisciplinaridade entre áreas da Linguística (dialetologia, lexicografia, sintaxe)

(SFRH/BD/27648/2006)
 (Cf. <http://www.clul.ul.pt/files/diadia/GloDiP.pdf>)

 CLUL

2016-09-17 1º ELI 9

CasualConc
Current File/Folder: alc_morf.txt

File Concord Cluster Collocation Word Count Corpus File Info

Search Span 60: 60: Sort Context

Context Word

KWIC	13 Found in	1 Files	File
1			Começa/VB-P-3S o/D cão/N a/P deixar/VB de/P comer/VB e/CONJ c alc_morf.txt
2			Começa/VB-P-3S o/D cão/N a/P deixar/VB de/P comer/VB e/CONJ começa/VB-P-3S a/P mirar-se/VB+SE //, e/CONJ <alt> </alt> mo alc_morf.txt
3			<inq> INQ2 Olhe, quando o cão começa a babar-se todo?... </inq> alc_morf.txt
4			is/ADV da/P+D-F azeitona/N //, está/ET-P-3S verde/ADJ-G //, começa-se/VB-P-3S+SE a/P fazer/VB preta/ADJ-F ./ alc_morf.txt
5			Quando/WADV começa/VB-P-3S a/P rebentar/VB a/D-F árvore/N ?/ alc_morf.txt
6			<in> INF </in> O/D que/WPRO começa/VB-P-3S lá/ADV a/P nascer/VB <break> (...)</break> 6/ alc_morf.txt
7			xo/N //, e/CONJ depois/ADV vão/VB-P-3P lá/ADV acima/ADV //, começam/VB-P-3P a/P baterem/VB-P-3P para/P baixo/N ./ alc_morf.txt
8			untar. Tem uma oliveira e no pé da oliveira, aqui em baixo, começam a nascer... </inq> alc_morf.txt
9			anos lá outra vez ao princípio. O senhor tem uma oliveira e começam, no pé da oliveira, começam a nascer assim umas coisa alc_morf.txt
10			se/CONJS elas/PRO vêm/VB-P-3P lá/ADV outro/OUTRO ovo/N //, começam/VB-P-3P a/P picar/VB e/CONJ partem/VB-P-3P os/D-P ovo alc_morf.txt
11			io. O senhor tem uma oliveira e começam, no pé da oliveira, começam a nascer assim umas coisas que o senhor tem que lá ir alc_morf.txt
12			<inq> INQ1 Pronto! Então podemos começar. Ó senhor Anselmo, quando o senhor </inq> alc_morf.txt
13			E/CONJ depois/ADV eles/PRO já/FP começavam/VB-D-3P <break> (...)</break> a/P roer/VB aquilo/D alc_morf.txt



2016-09-17 1º ELI 10

FileMaker Pro Advanced - [verbo_1tese]

File Edit View Insert Format Records Scripts Tools Window Help

----- GloDiP -----

verbo ler SE Dial

INAC I T T2 T3 DIT COP T.PRED INF IMP PP PC

1. percorrer com a vista e conhecer letras, reunindo-as em palavras.

OBS

[+SN] SE PRED SE2 dial2 flex

N

A minha tia lia aquelas histórias. C-GRJ06; C

como um que ande na escola e que lê o jornal e que lê isto e que lê aquilo e que não erra. S-PAL16; Eu tenho lido vários livros". S-ALV47; S

Eu já li lá os estatutos que ele o baldio tem uns estatutos. A-CRV51; A

M

OBS2



2016-09-17 1º ELI 11

começar v. INF, T, INAC, INACIT, IMPIT2, COP, MP
INF
1. iniciar a fazer alguma coisa.
[+a]
 Começaram a aparecer outros agasalhos e as pessoas começaram a ir... A-CDR07; Um dia estava para cear e eu estava assentado e ela lá a dar-me a ceia, começou ela a dizer: C-COV02; Ele em Março, começa a gente a tratar da terra: C-MST31; O pêlo era todo cortado, primeiro, para então depois começar a continuar a acartar vinho. M-PST06; E depois, começou-se assim a constar e foi aos ouvidos aos pais dela. N-PFT25
[+a] <flex>
 depois começou a aparecer os colchões Molaflex, que já nem se usa colchões de riscado. A-CDR07; E depois já começa as árvores a rebentar S-AAL30; estendem uns plásticos por baixo, e depois vão lá acima, começam a baterem para baixo. S-ALC17
[+de] <●>
 Agora semeados em Janeiro, quando é ele para meados de Abril, é que eles começam de aumentar. A-CRV55; Já algum dia, quando me começa de lembrar, não havia linhas a vender como há agora nas... C-PVC08; depois quando ele começa de abrir assim uns gricheirinhos, N-PFT08

 CLUL

2016-09-17 1º ELI 12

Grupo de Dialetologia e Diacronia do CLUL

Corpus CORDIAL-SIN (1999-...)
 Transcrição
 Anotação POS (*Part-of-Speech*)
 Anotação sintática (em curso)



GloDiP (2007-2011)
 Glossário dos dialetos portugueses

Projeto WOChWEL (2012-...)
Anotação POS
Anotação sintática

 CLUL

2016-09-17

1º ELI 13

WOCHWEL: Word Order and Word Order Change in Western European Languages

- ◆ anotação POS
- ◆ anotação sintática
- textos medievais (séc. XIII)
 - ✓ Textos do ciclo arturiano (c. 360.000 palavras)
 - ✓ Livro de José de Arimateia (Castro 1984)
 - ✓ Demanda do Santo Graal (Toledo 2012-2015)
 - ✓ Textos notariais (c. 80.000 palavras, em curso)

(Cf. <http://alfclul.clul.ul.pt/wochwel/index.html>)



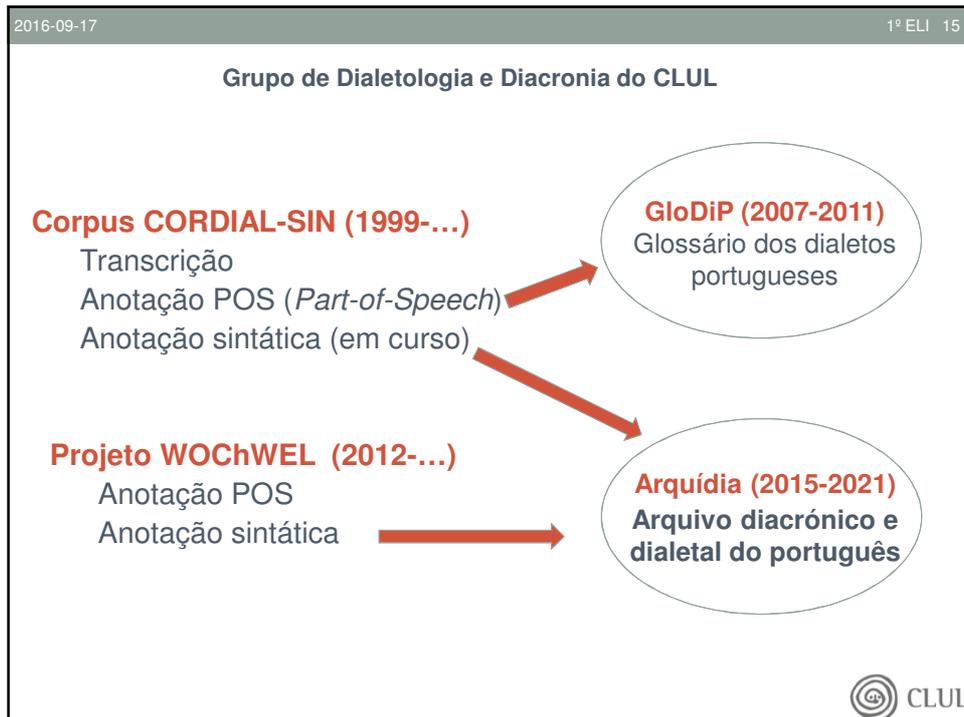
2016-09-17

1º ELI 14

“Syntactically parsed treebanks are even more useful than POS tagged corpora in linguistic research, as they not only provide part-of-speech information for individual words but also indicate constituent types and membership”

(McEnery et al 2006)





2016-09-17 1º ELI 16

corpora do português que seguem a anotação dos Penn corpora

- WOChWEL (séc. XIII)
- **Arquídia (séc. XIV):** *Crónica Geral de Espanha* (90.000 palavras, em curso)
- Tycho Brahe (séc. XV – XX)
- Post-Scriptum (séc. XVI – XX)
- CORDIAL-SIN (séc. XX)

CLUL

2016-09-17 1º ELI 17

- Construir uma ferramenta que permita pesquisas online e reúna as aceções e construções gramaticais para uma determinada palavra em diferentes períodos da história da língua e diferentes variedades dialetais;
- Ampliar recursos já existentes (anotação de textos medievais e dialetais) permitindo que as amostras sejam representativas;
- Disponibilizar os novos textos anotados;
- Explorar o potencial lexicográfico dos corpora anotados sintaticamente (informação sintática mais completa; os exemplos mais facilmente extraídos e mais adequadamente escolhidos);
- Alargar o conhecimento do léxico e da sintaxe do português, cruzando informação histórica e dialetal;
- Refinar ou inovar os sistemas de anotação testando-os com os dados de diferentes variedades do português, no âmbito do sistema Penn Treebank.

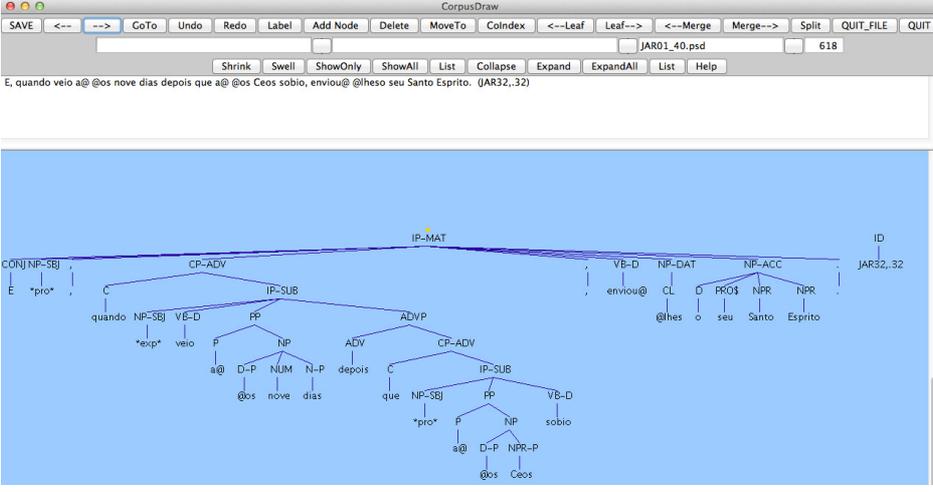
(SFRH/BPD/99678/2014)



2016-09-17 1º ELI 18

CorpusDraw

(parte integrante do *CorpusSearch*: <http://corpussearch.sourceforge.net>)



The screenshot shows the CorpusDraw application window with a menu bar (SAVE, GoTo, Undo, Redo, Label, Add Node, Delete, MoveTo, Colindex, Leaf, Merge, Split, QUIT_FILE, QUIT) and a toolbar (Shrink, Swell, ShowOnly, ShowAll, List, Collapse, Expand, ExpandAll, List, Help). The main area displays a syntax tree for the sentence: "E, quando veio a @os nove dias depois que a @os Ceos sobio, enviou @lheseo seu Santo Esprito. (JAR32,,32)". The tree is rooted at IP-MAT and branches into various syntactic categories like CP-ADV, IP-SUB, ADVP, and NP-ACC, with terminal nodes corresponding to the words in the sentence.



2016-09-17 1º ELI 19

Editor de texto

```
( (IP-MAT (CONJ E)
  (NP-SBJ *pro*)
  (, ,)
  (CP-ADV (C quando)
    (IP-SUB (NP-SBJ *exp*)
      (VB-D veio)
      (PP (P a@)
        (NP (D-P @os) (NUM nove) (N-P dias)))
      (ADVP (ADV depois)
        (CP-ADV (C que)
          (IP-SUB (NP-SBJ *pro*)
            (PP (P a@)
              (NP (D-P @os) (NPR-P Ceos)))
            (VB-D sobio))))))
  (, ,)
  (VB-D enviou@)
  (NP-DAT (CL @lhes))
  (NP-ACC (D o) (PRO$ seu) (NPR Santo) (NPR Esprito))
  (. .))
(ID JAR32,.32))
```

CLUL

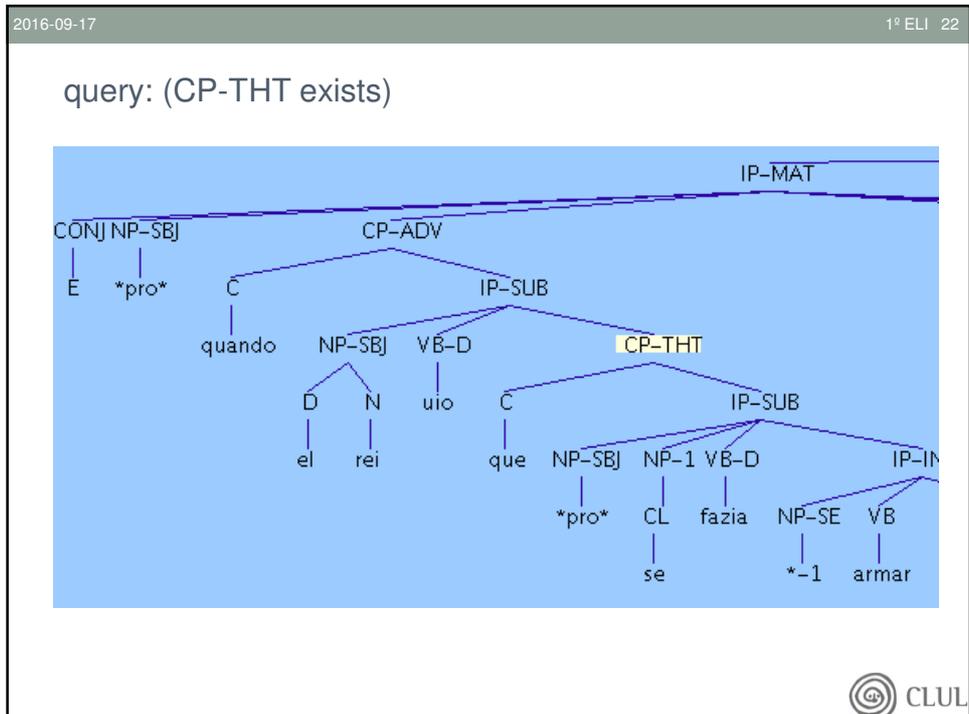
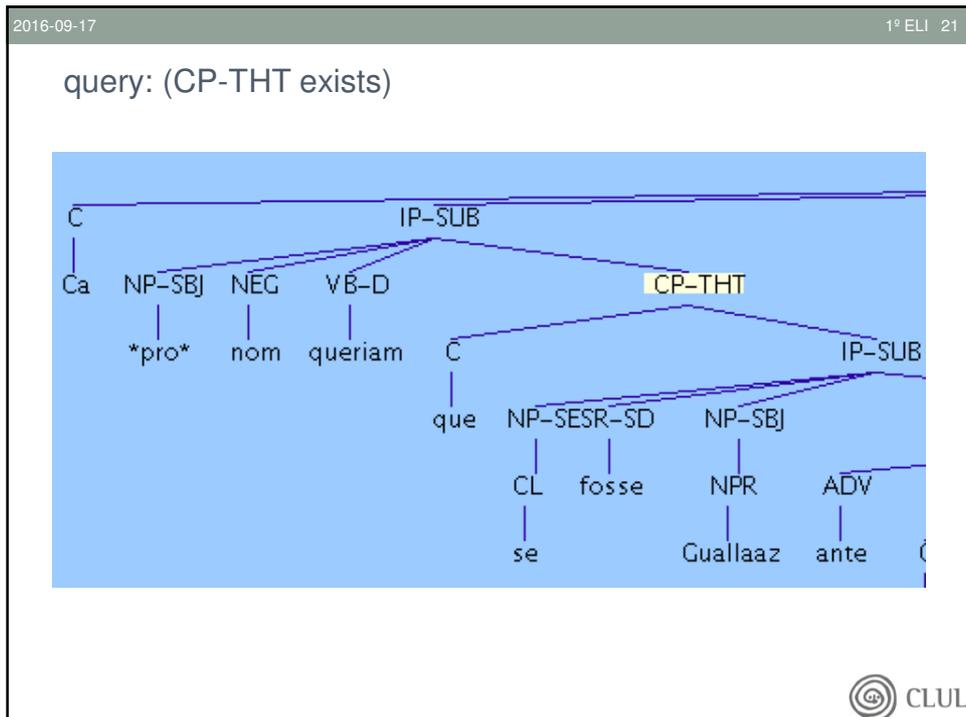
2016-09-17 1º ELI 20

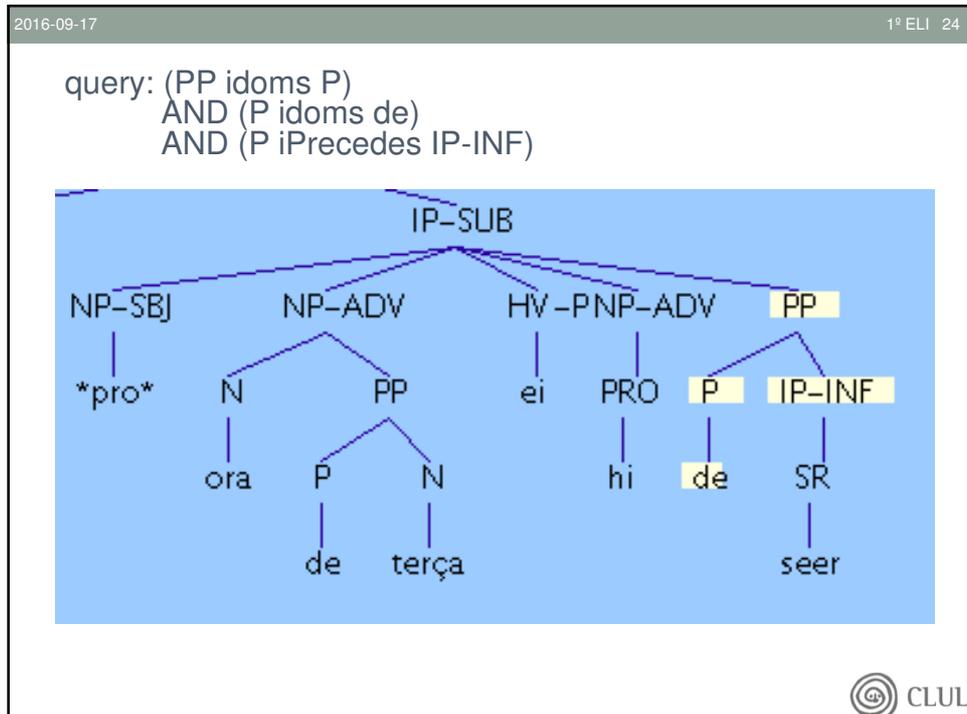
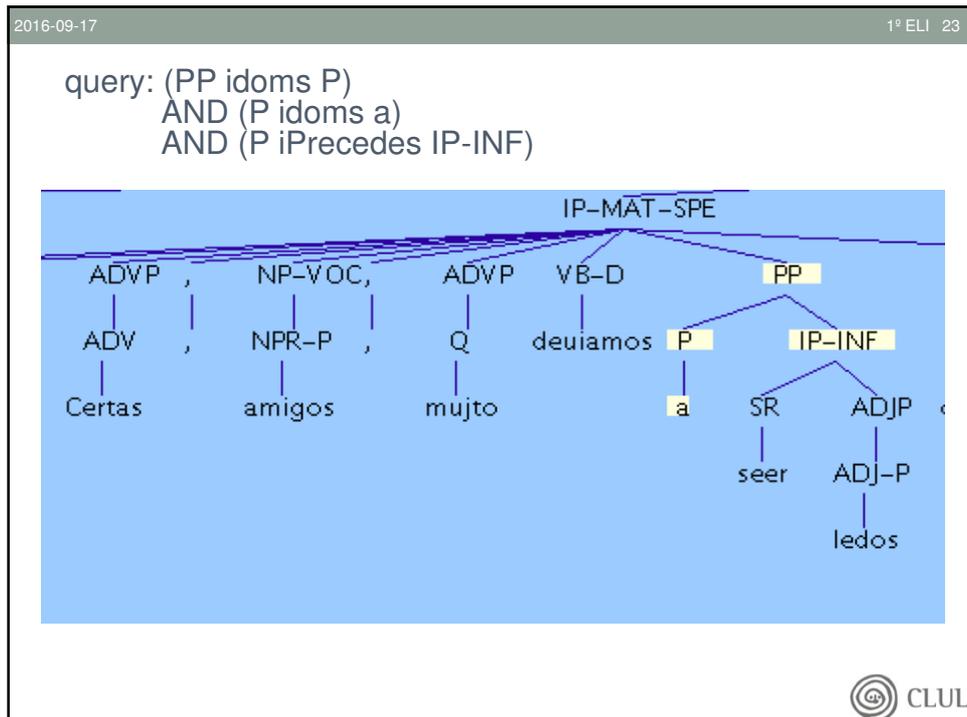
- **CorpusSearch:**
um motor de busca concebido para pesquisa e revisão / construção de *corpora* linguísticos que seguem a anotação dos *Penn Corpora* (Cf. Randall 2007)

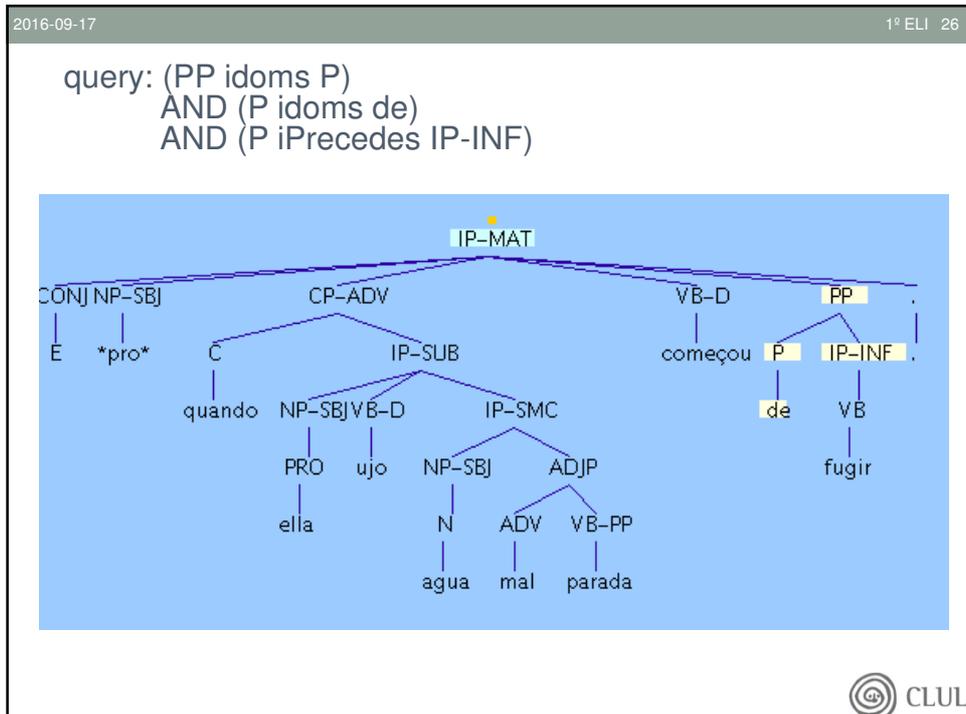
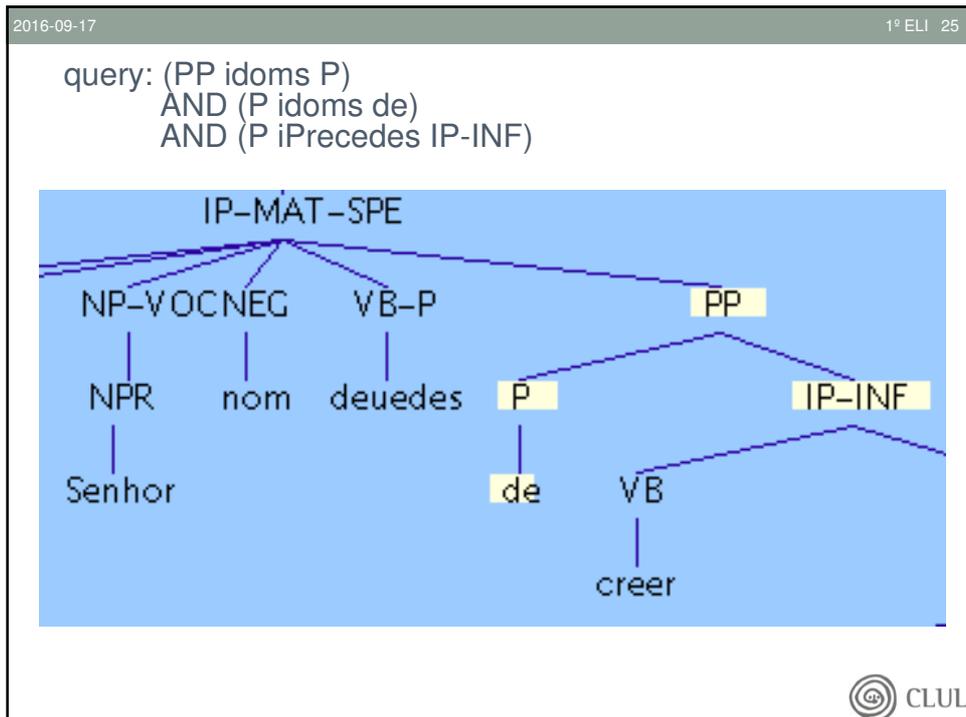
↓

queries

CLUL







2016-09-17

1º ELI 27

**começar v. INF, T, INAC, INAC|T, IMP|T2, COP, MP
INF**

1. iniciar a fazer alguma coisa.

[+a]

*El-rey **começou a pensar** muito DSG13*

E, dally adyâte, *começou o conde a manter sua terra mais seguramente que ante fazia.* CGE320

*Começaram a aparecer outros agasalhos e as pessoas começaram a ir... A-CDR07; Um dia estava para cear e eu estava assentado e ela lá a dar-me a ceia, *começou* ela a dizer: C-COV02; Ele em Março, *começa* a gente a tratar da terra: C-MST31; O pêlo era todo cortado, primeiro, para então depois *começar a continuar* a acartar vinho. M-PST06; E depois, *começou-se* assim a constar e foi aos ouvidos aos pais dela. N-PFT25*



2016-09-17

1º ELI 28

**começar v. INF, T, INAC, INAC|T, IMP|T2, COP, MP
INF**

1. iniciar a fazer alguma coisa.

[+de] < ● >

*E quando ella uio agua mal parada, **começou de fugir** DSG13*

E quando forõ e se *começaron de ferir* veria des aos primeiros golpes cayr muytos mouros... CGE329,1.23

*Agora sementeados em Janeiro, quando é ele para meados de Abril, é que eles *começam de aumentar*. A-CRV55; Já algum dia, quando me *começa de lembrar*, não havia linhas a vender como há agora nas... C-PVC08; depois quando ele *começa de abrir* assim uns gricheirinhos, N-PFT08*



2016-09-17

1º ELI 29

dever v. INF, DIT, T**INF**

1. ter a obrigação ou a responsabilidade de.

[+de] < ☺ >Senhor nom *deuedes de creer...* DSG13**E tal senhor como este *devemos de amar e servyr e poer* ã elle os nossos corações e os nossos feitos ... CGE341**

E ele dizia muita vez que nós que não sabíamos comer fruta, que a gente nunca *devia de debulhar* a fruta. C-MTV47; não lhe dão o fermento que *deve de ser*, M-PST16; Só cria aquele que *deve de criar*, o resto vende-o pequeno... N-LAR12; "Não *devia de mandar* os rapazes, que eu não quero que os rapazes mexam em nada de ninguém. S-STJ72



2016-09-17

1º ELI 30

dever v. INF, DIT, T**INF**

1. ter a obrigação ou a responsabilidade de.

[+a] < ☺ >Certas amigos, mujto *deuiamos a ser ledos...* DSG13**Ca este Senhor nõ *devemos a leixar por a semelhança da besta* , que nõ he ne hũa cousa , CGE341**

o golpe *deve-se a fazer* sempre assente dentro duma norma para não ficar nem alto demais, nem baixo demais. S-CBV03



2016-09-17 1º ELI 31

sair v. INAC|T, INAC, INAC|T2, T, T2, I, COP
INAC|T
 1. passar de dentro para fora; ir para fora, ou deixar um lugar; aparecer em público, ir à rua.
[de+SN]
Como o conde dom Fernã Gonçalvez *sayu da prisom e das razões da condessa com el rey* . CGE354
 No dia do bodo, então, o Senhor Espírito Santo quando *sai daqui da igreja*, que vem de casa do imperador coroadado, A-TRC19; Eu nunca *saí daqui*. C-GRJ05
[de+SN] <SE> ←
E *sairon-sse da villa muy emcubertamente* . CGE354
[a+INF] ←
E , quando chegarõ apreto da villa , *sairõ-no a receber cõ muy grandes procissões muyto honrradamente* . CGE348

entrar v. INAC|T, INAC, INAC|T2, T
T ←
 19. ir para dentro.
[+SN/Acus] ←
E *desi foram logo a Jeen e matarom quantos Berberiiis hi acharom; e foram logo a Murça e entrarõ-na*. CGE411

 CLUL

2016-09-17 1º ELI 32

Arquídia:

- anotação de textos (séc. XIV; dialetais)
- recurso online
 - acesso a pesquisas simples: entrada (informação lexicográfica + informação sintática)
 - acesso a pesquisas avançadas: construções sintáticas

Recurso para o estudo da sintaxe e para a lexicografia

 CLUL

OBRIGADA!



Referências

BIKEL, Dan (2004): *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. Diss. PhD. University of Pennsylvania. **CASTRO**, Ivo (1984): *Livro de José de Arimateia* (Estudo e Edição do COD. ANTT 643). Dissertação de Doutoramento. Universidade Lisboa. **GALVES**, C. & **P. FARIA**. (2010). *Tycho Brahe Parsed Corpus of Historical Portuguese*, <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html>; **KEPLER**, Fábio (2005): *Um etiquetador morfo-sintático baseado em cadeias de Markov de tamanho variável*. Dissertação de Mestrado. São Paulo: Instituto de Matemática e Estatística da Universidade de São Paulo. **MARQUILHAS**, R. (Coord.). 2014. *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*, <http://ps.clul.ul.pt>. **MARTINS**, A. M. (Coord.) (2012-2015). *Word order and word order change in western European languages*, <http://alfclul.clul.ul.pt/wochwel/> **MARTINS**, A. M. (Coord.) (1999-...). *Syntax-oriented corpus of Portuguese dialects*, <http://www.clul.ul.pt/en/resources/212-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects>. **MARTINS**, Ana Maria (1994): *Clíticos na História do Português*. Lisboa: Universidade de Lisboa. Tese de doutoramento. **RANDALL**, B. (2007). *CorpusSearch 2*, <http://corpussearch.sourceforge.net>; **TOLEDO NETO**, Sílvia (2012-2015): *Transcrição / Edição da Demanda do Santo Graal*. Manuscrito não publicado.

