ARQUÍDIA

Um recurso em construção

Sandra Pereira
spereira@clul.ul.pt

V SIMELP – Simpósio Mundial de Estudos de Língua Portuguesa
Simpósio 41: Dicionarística Portuguesa: Investigação e Projetos em Curso
8-11 outubro, Lecce

Grupo de Dialetologia e Diacronia do CLUL

- Corpus CORDIAL-SIN (1999-…)
  - Anotação POS (*Part-of-Speech*)
  - Anotação sintática (em curso)
- ✓GloDiP: Glossário dos Dialetos Portugueses com Informação Sintática (SFRH/BD/27648/2006)
- Projeto WOChWEL (2012-2015)
  - Anotação POS
  - Anotação sintática
- ➢Arquídia: Arquivo diacrónico e dialetal do português (SFRH/BPD/99678/2014)

CLUL

- corpus: CORDIAL-SIN (*Corpus Dialetal para o Estudo da Sintaxe*)
- anotação: POS

(1) Estendem (...) uns sacos por baixo – agora é de plástico -, estendem uns plásticos por baixo, e depois vão lá acima, começam a baterem para baixo. <ALC01>

(1') 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 ./. <ALC01>

(Cf. http://www.clul.ul.pt/en/research-teams/212-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects)

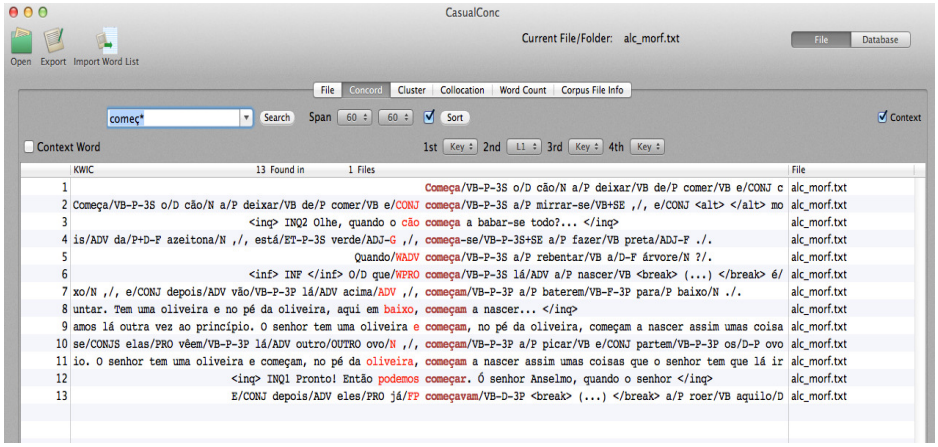
CLUL

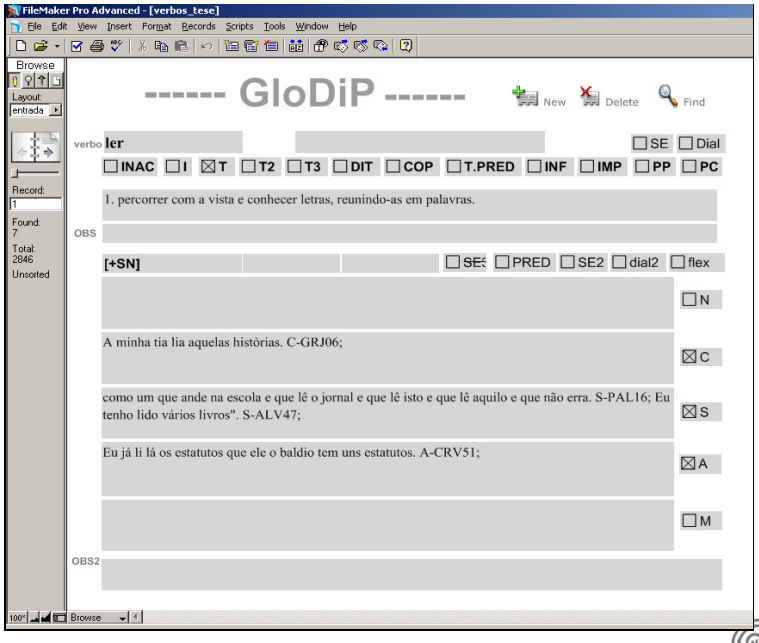
GloDiP (Pereira 2012)

- 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 monolingue baseada em *corpora*)
- Explorar as potencialidades de um *corpus* dialetal
- Contribuir para a interdisciplinaridade entre áreas da Linguística (dialetologia, lexicografia, sintaxe)

(Cf. http://www.clul.ul.pt/files/diadia/GloDiP.pdf)

CLUL


CasualConc
Open Export Import Word List
Current File/Folder: alc_morf.txt
File Database
File Concord Cluster Collocation Word Count Corpus File Info
começ* Search Span 60 60 Sort Context
Context Word
1st Key 2nd L1 3rd Key 4th Key
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 mirrar-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 <inf> INF </inf> O/D que/WPRO começa/VB-P-3S lá/ADV a/P nascer/VB <break> (...) </break> é/ 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-F-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 amos 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 vêem/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


FileMaker Pro Advanced - [verbos_tese]
File Edit View Insert Format Records Scripts Tools Window Help
Browse
Layout: entrada
Record: 1
Found: 7
Total: 2846
Unsorted
------ GloDiP ------
New Delete Find
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

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

WOChWEL (*Word Order and Word Order Change in Western European Languages*)

- anotação POS
- anotação sintática
- textos medievais (séc. XIII)
  - ✓ Textos do ciclo arturiano
  - ✓ Textos notariais

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

CLUL

"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"

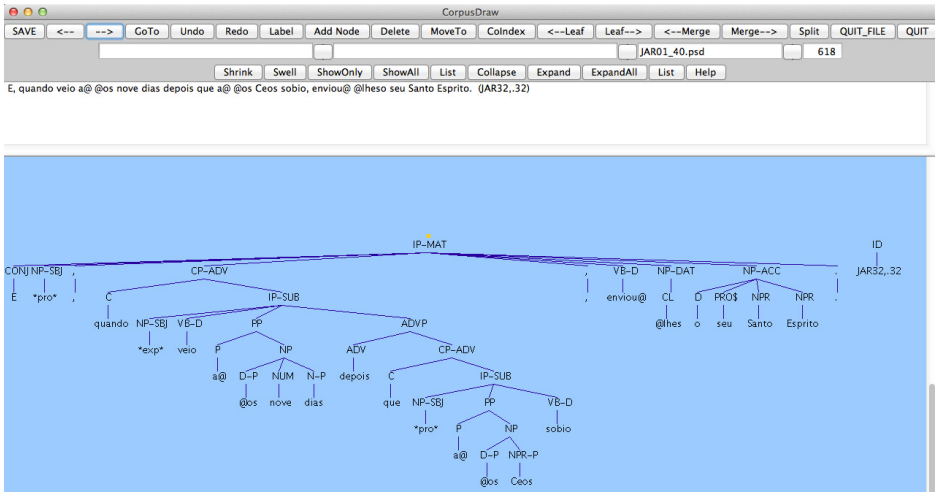
(McEnery et al 2006)

Anotação Penn *corpora*:

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

CorpusDraw

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



Editor de texto

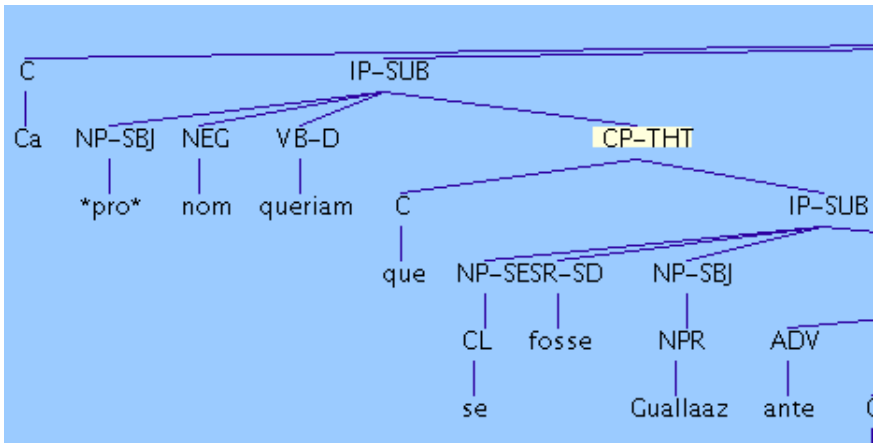
```
( (IP-MAT (CONJ E)
        (NP-SBJ *pro*)
        (, ,)
        (CP-ADV (C quando)
              (IP-SUB (NP-SBJ *exp*)
                    (VB-D veio)
                    (PP (P a@)
                        (NP (D-P @os) (NUM nove) (N-P dias)))
                    (ADVP (ADV depois)
                        (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))
```

- ***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

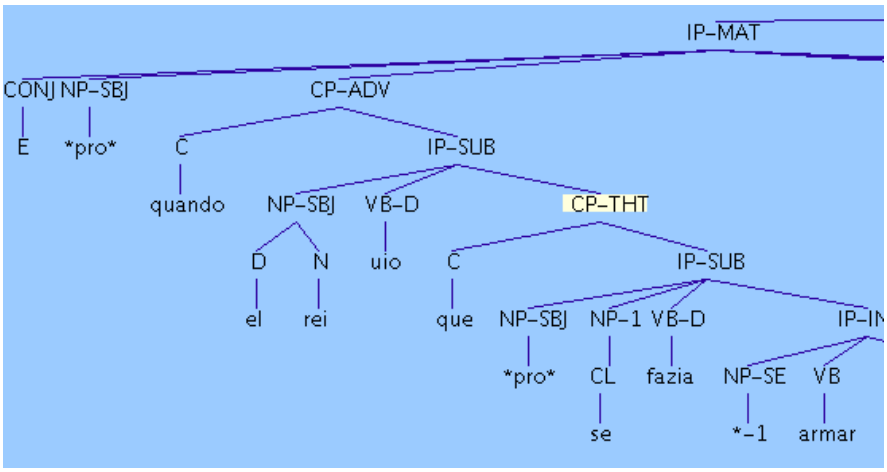
CLUL

query: (CP-THT exists)

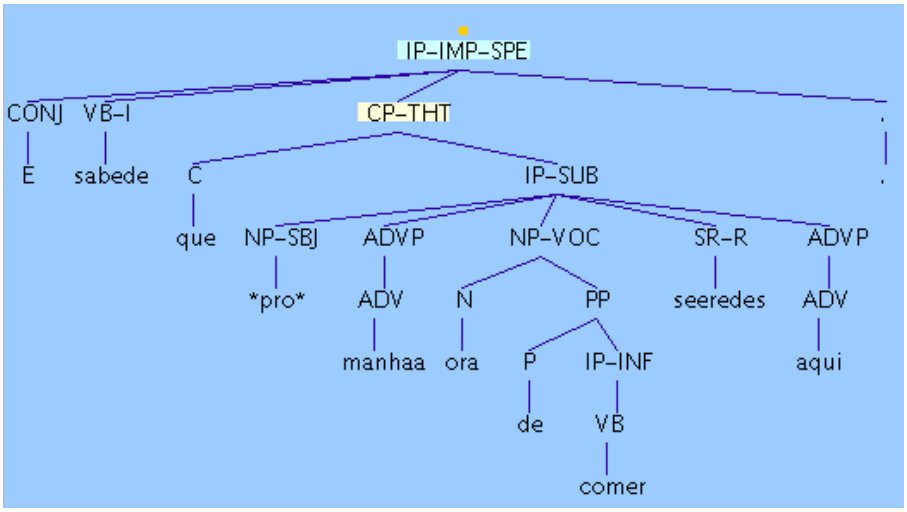


CLUL

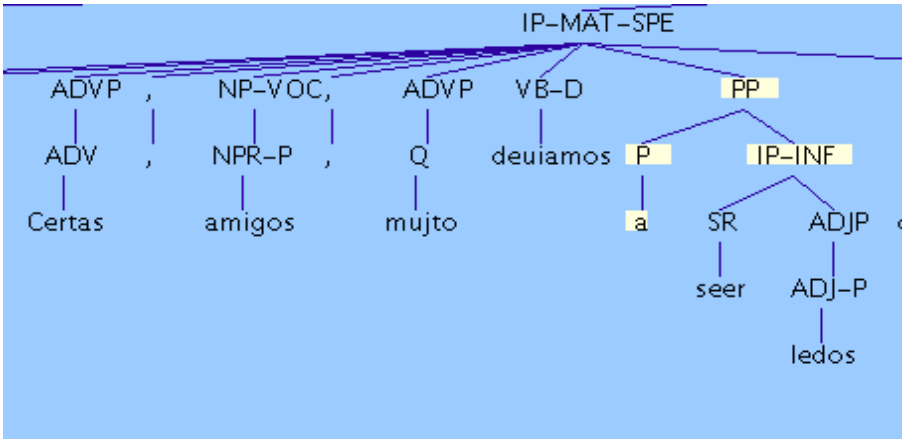
query: (CP-THT exists)



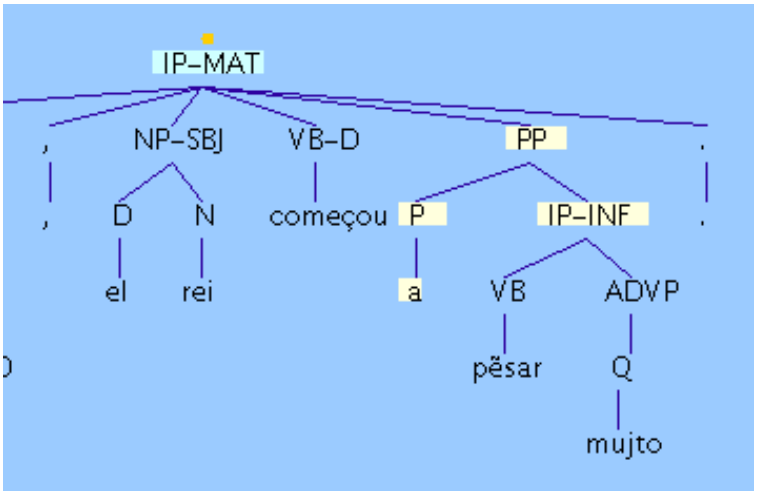
query: (CP-THT exists)



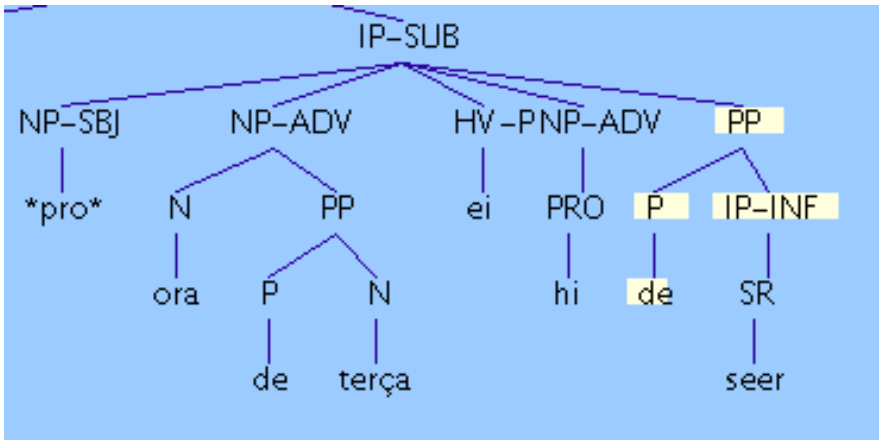
query: (PP idoms P)
AND (P idoms a)
AND (P iPrecedes IP-INF)



query: (PP idoms P)
AND (P idoms a)
AND (P iPrecedes IP-INF)

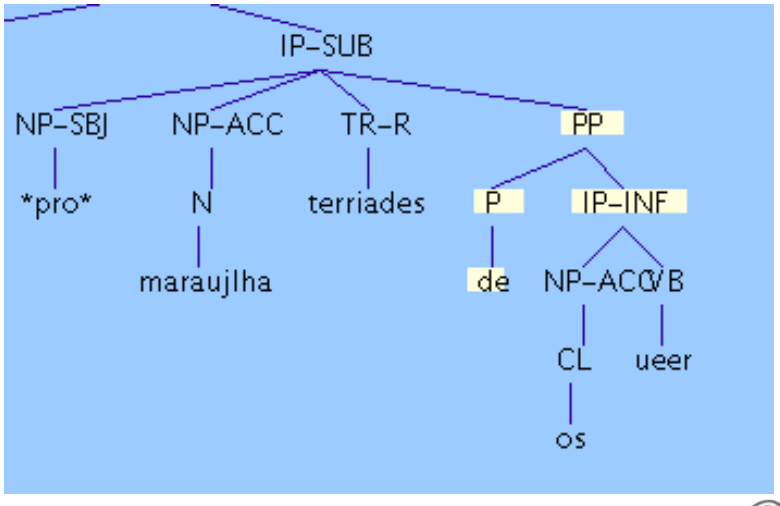


query: (PP idoms P)
AND (P idoms de)
AND (P iPrecedes IP-INF)



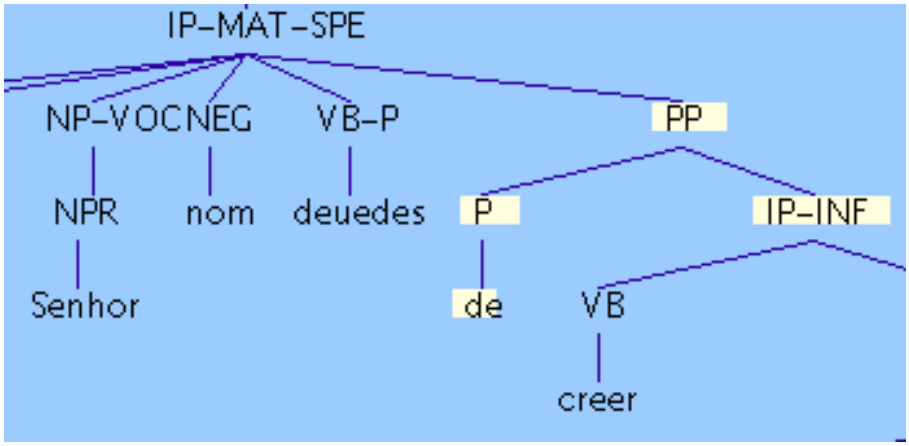
CLUL

query: (PP idoms P)
AND (P idoms de)
AND (P iPrecedes IP-INF)

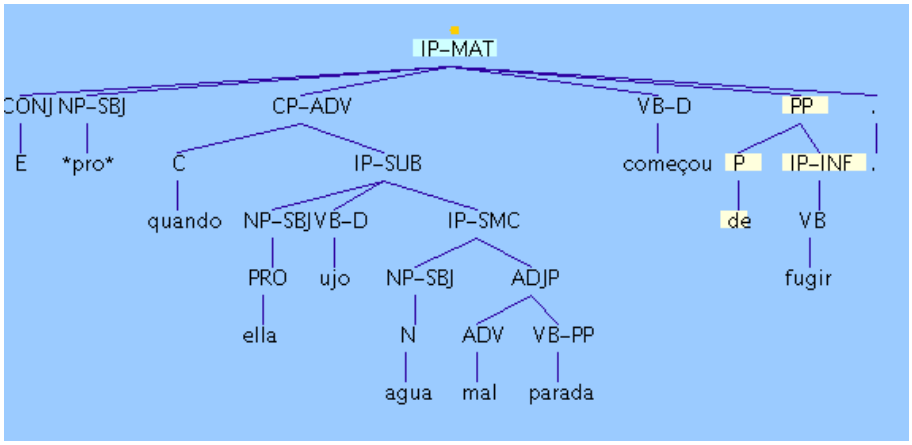


CLUL

query: (PP idoms P)
AND (P idoms de)
AND (P iPrecedes IP-INF)



query: (PP idoms P)
AND (P idoms de)
AND (P iPrecedes IP-INF)



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

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

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

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

dever *v.* **INF, DIT, T**
INF
1. ter a obrigação ou a responsabilidade de.
[+de] < ☻ >
Senhor nom deuedes de creer... DSG13
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

dever *v.* **INF, DIT, T**
INF
1. ter a obrigação ou a responsabilidade de.
[+a] < ☻ >
Certas amigos, mujto deuiamos a ser ledos... DSG13
o golpe deve-se a fazer sempre assente dentro duma norma para não ficar nem alto demais, nem baixo demais. S-CBV03

(2) Quando o dia começou a esclarecer … DSG420,1.1/ID

(3) E, quando chegarõ aí, nõ nos saluarõ . DSG425,1.20/ID

(4) Nom quedou de chorar porque ujo ca se auja de partir delle. DSG05,1.5/ID (também “estar quedo”)

(5) Entõ esteuerõ e atenderõ ata que chegou Galaaz a eles. DSG488,1.34/ID

(6) E tanto andou ates que chegou a hũũ valle … DSG220,1.4/ID

(7) E Brioberis disse outrosi desi como aquel que mais queria seu enxeco ca sa paz, DSG401,1.33/ID

CLUL

Arquídia:

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

Recurso para o estudo da sintaxe e para a lexicografia

CLUL

OBRIGADA!

CLUL

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 cadeiasi 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**, Sílvio (2012-2015): Transcrição / Edição da Demanda do Santo Graal. Manuscrito não publicado.

CLUL