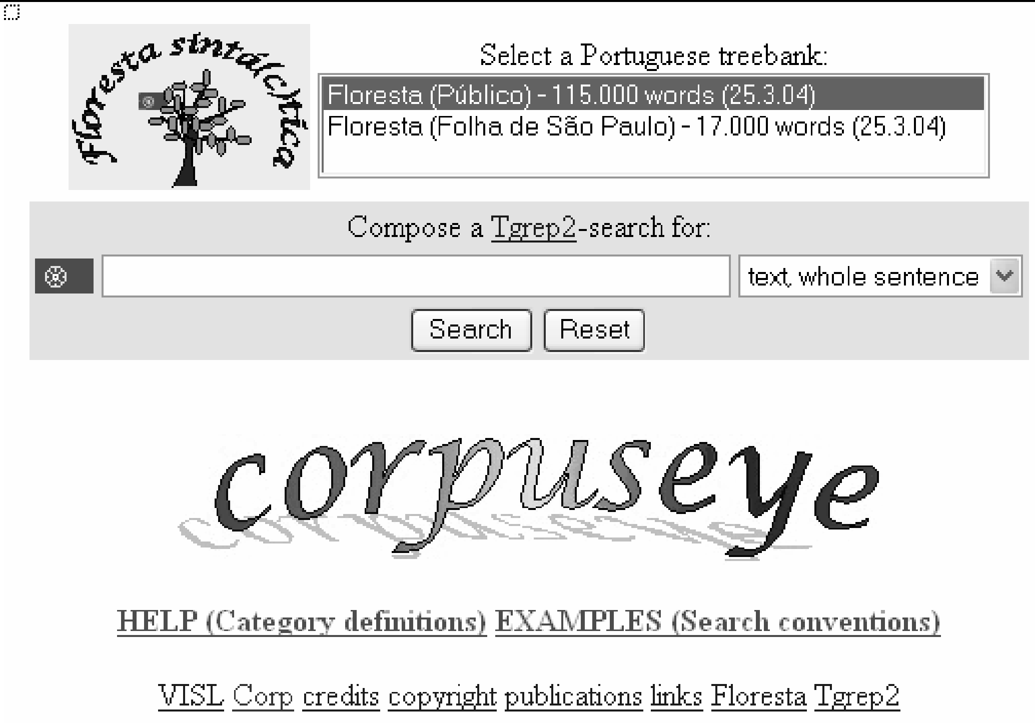
Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface

Eckhard Bick
University of Southern Denmark



Select a Portuguese treebank:

Floresta (Público) - 115.000 words (25.3.04)

Floresta (Folha de São Paulo) - 17.000 words (25.3.04)

Compose a Tgrep2-search for:

text, whole sentence

Floresta sintá(c)tica
Search
Reset
corpuseye
HELP (Category definitions) EXAMPLES (Search conventions)
VISL Corp credits copyright publications links Floresta Tgrep2

Introduction

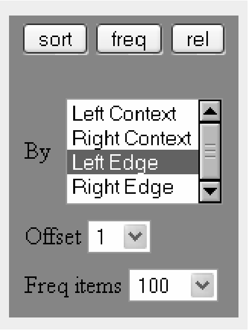
The CorpusEye project (http://corp.hum.sdu.dk) at the University of Denmark aims at designing and programming an internet based corpus search interface that (1) offers standardised search tools and a unified descriptive formalism across different corpus types and different languages, and (2) allows users to exploit grammatical information in annotated corpora in a user-friendly and menu-based way. All corpora in CorpusEye have been annotated with VISL's Constraint Grammar based parsers, in the case of treebanks using an additional PSG module or equivalent (Bick 2003). At the time of writing, the material covers 8 languages and ca. 600 million words.

CorpusEye's internal search database uses the IMS' Corpus Query Protocol (Christ 1994, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/) for CG-korpora and the linux-tool tgrep2 (http://tedlab.mit.edu/~dr/Tgrep2/) for the treebanks[1].

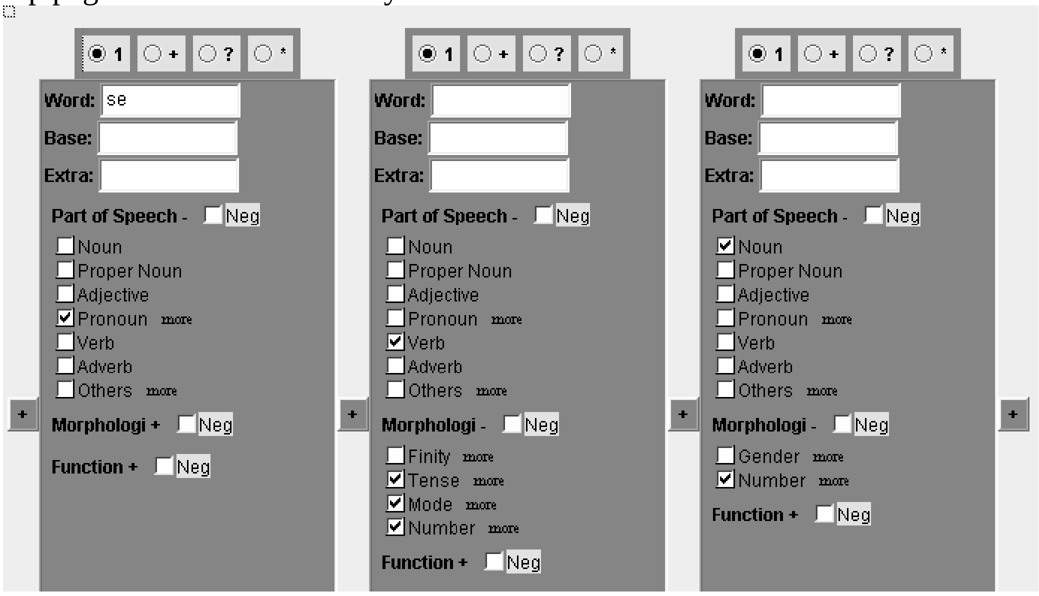
The menu-based CQP interface: Word based form and function

Though CorpusEye allows direct use of CQP-speak and so-called regular expressions (joker characters, sets and mathematical operators), the project primarily targets the humanistic user without prior knowledge of the formal aspects of corpus linguistics (teachers, literary researchers, lexicographers etc.). Thus, it is possible to get started with simple text searches, presented in concordance format. With a single click, the user can produce statistical overviews (absolute or relative frequencies) for words in a given position of the search string or its context.

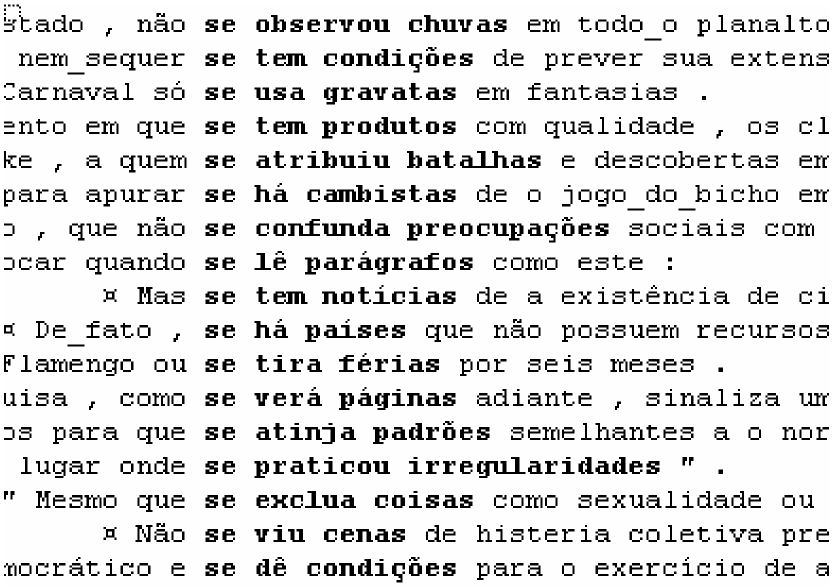
[1] Internet use of CQP was inspired by other search interfaces programmed earlier by Paul Meurer for Norwegian (Oslo University, http://www.hf.uio.no/tekstlab/) and Diana Santos for Portuguese (Linguateca, http://www.linguateca.pt). The CQP interface described here was designed by Eckhard Bick, with substantial programming help by Poul Henriksen and Nikolaj H. Nielsen.



In the graphical interface, it is possible to enter lexemes (base forms) for lemmatised searches, or to select part of speech (noun, verb etc.), morphology (singular, present tense etc.) or syntactic function (subject, direct object etc.) from a word-linked menues. Choices can be negated, and search fields can be marked for optionality or repetition. In the end, all search specifics will be translated - invisibly for the user - into a CQP search expression. Most buttons in the interface are self-explanatory through popup-windows, and the top page offers an introductory Flash film.

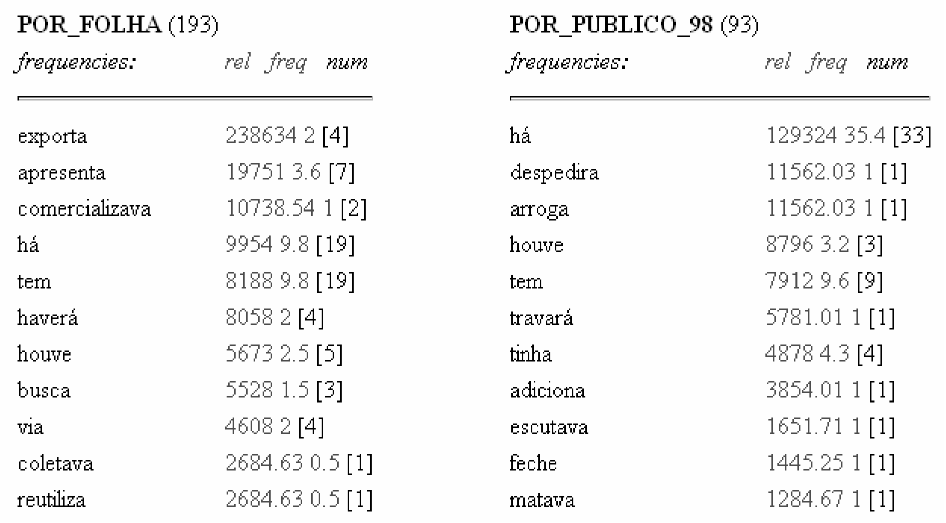


In the example, we search for the pronoun "se" (excluding the conjunction "se"!), followed by a finite verb in the singular and then a noun in the plural. Note that the illustration doesn't show the unfolded tense, mode and number fields hiding, in the screendump, the specific choices made. The resulting concordance exemplifies the use of a singular, impersonal "se" with a "subject-like" function - proven by the number-inflexion mismatch between its verb and the supposed reflexive noun-base (now interpretable as "object").



```
stado , não se observou chuvas em todo_o planalto
 nem_sequer se tem condições de prever sua extens
Carnaval só se usa gravatas em fantasias .
ento em que se tem produtos com qualidade , os cl
ke , a quem se atribuiu batalhas e descobertas em
para apurar se há cambistas de o jogo_do_bicho em
o , que não se confunda preocupações sociais com
ocar quando se lê parágrafos como este :
      ¤ Mas se tem notícias de a existência de ci
¤ De_fato , se há países que não possuem recursos
Flamengo ou se tira férias por seis meses .
uisa , como se verá páginas adiante , sinaliza um
os para que se atinja padrões semelhantes a o nor
 lugar onde se praticou irregularidades " .
" Mesmo que se exclua coisas como sexualidade ou
      ¤ Não se viu cenas de histeria coletiva pre
mocrático e se dê condições para o exercício de a
```

A relative frequency analysis on the verb (position-defined as left edge of search string + 1) can be run in order to examine which verbs are more likely to appear with impersonal "se" than they would otherwise be in running text (normalised frequencies). Note that the statistics can be done for 2 corpora at a time, comparatively. Here, European Portuguese (Público) is compared with Brazilian Portuguese (Folha de São Paulo).



POR_FOLHA (193)

frequencies:	*rel*	*freq*	*num*
exporta	238634	2	[4]
apresenta	19751	3.6	[7]
comercializava	10738.54	1	[2]
há	9954	9.8	[19]
tem	8188	9.8	[19]
haverá	8058	2	[4]
houve	5673	2.5	[5]
busca	5528	1.5	[3]
via	4608	2	[4]
coletava	2684.63	0.5	[1]
reutiliza	2684.63	0.5	[1]

POR_PUBLICO_98 (93)

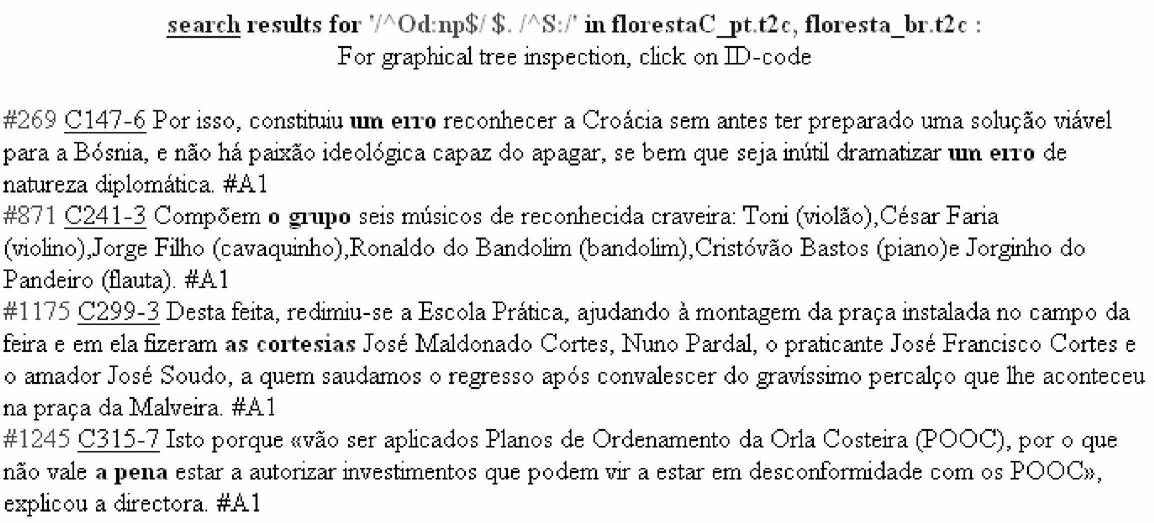
frequencies:	*rel*	*freq*	*num*
há	129324	35.4	[33]
despedira	11562.03	1	[1]
arroga	11562.03	1	[1]
houve	8796	3.2	[3]
tem	7912	9.6	[9]
travará	5781.01	1	[1]
tinha	4878	4.3	[4]
adiciona	3854.01	1	[1]
escutava	1651.71	1	[1]
feche	1445.25	1	[1]
matava	1284.67	1	[1]

Searching for constituents and constituent structure

The second interface, based on tgrep2, more directly targets the structural information contained in treebanks, allowing constituent searches rather than word based searches. Thus, when searching for an object followed by a subject, the search can be formulated as a 2-element search rather than having to provide for optional modifier positions around the object and subject heads:

```
/Od:np$/ $. /^S:np/
```

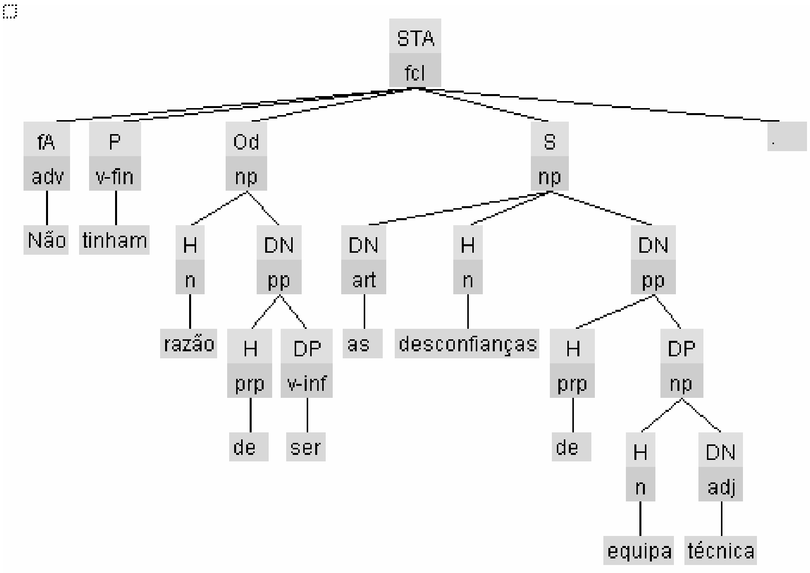
The convention used here, is VISL's form & function duality, with a function symbol in upper case and a form symbol in lower case, separated by a colon. In the search string, functions are Od (direct object) and S (subject), forms are np (noun phrase). // means a constituent (one ore more words), $. means "sisterhood" (same mother-node). In the example from the Público part of the Floresta Sintá(c)tica treebank, results are given in running text, but tree-structure or Penn-treebank style constituent bracketing can be chosen alternatively.



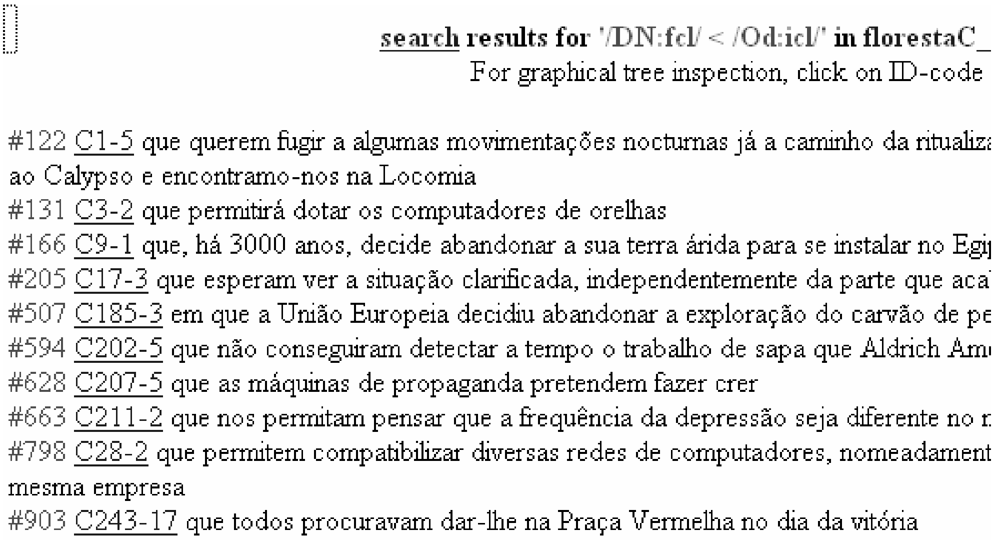
search results for '/^Od:np$/ $. /^S:/' **in floresta C_pt.t2c, floresta_br.t2c** :
For graphical tree inspection, click on ID-code

#269 C147-6 Por isso, constituiu **um erro** reconhecer a Croácia sem antes ter preparado uma solução viável para a Bósnia, e não há paixão ideológica capaz do apagar, se bem que seja inútil dramatizar **um erro** de natureza diplomática. #A1
#871 C241-3 Compõem **o grupo** seis músicos de reconhecida craveira: Toni (violão),César Faria (violino),Jorge Filho (cavaquinho),Ronaldo do Bandolim (bandolim),Cristóvão Bastos (piano)e Jorginho do Pandeiro (flauta). #A1
#1175 C299-3 Desta feita, redimiu-se a Escola Prática, ajudando à montagem da praça instalada no campo da feira e em ela fizeram **as cortesias** José Maldonado Cortes, Nuno Pardal, o praticante José Francisco Cortes e o amador José Soudo, a quem saudamos o regresso após convalescer do gravíssimo percalço que lhe aconteceu na praça da Malveira. #A1
#1245 C315-7 Isto porque «vão ser aplicados Planos de Ordenamento da Orla Costeira (POOC), por o que não vale **a pena** estar a autorizar investimentos que podem vir a estar em desconformidade com os POOC», explicou a directora. #A1

In the concordance, ID's are links to java-based graphical representations, which can be manipulated in various ways - unfolded layer by layer, complexity-filtered or even rebuilt interactively for teaching purposes.



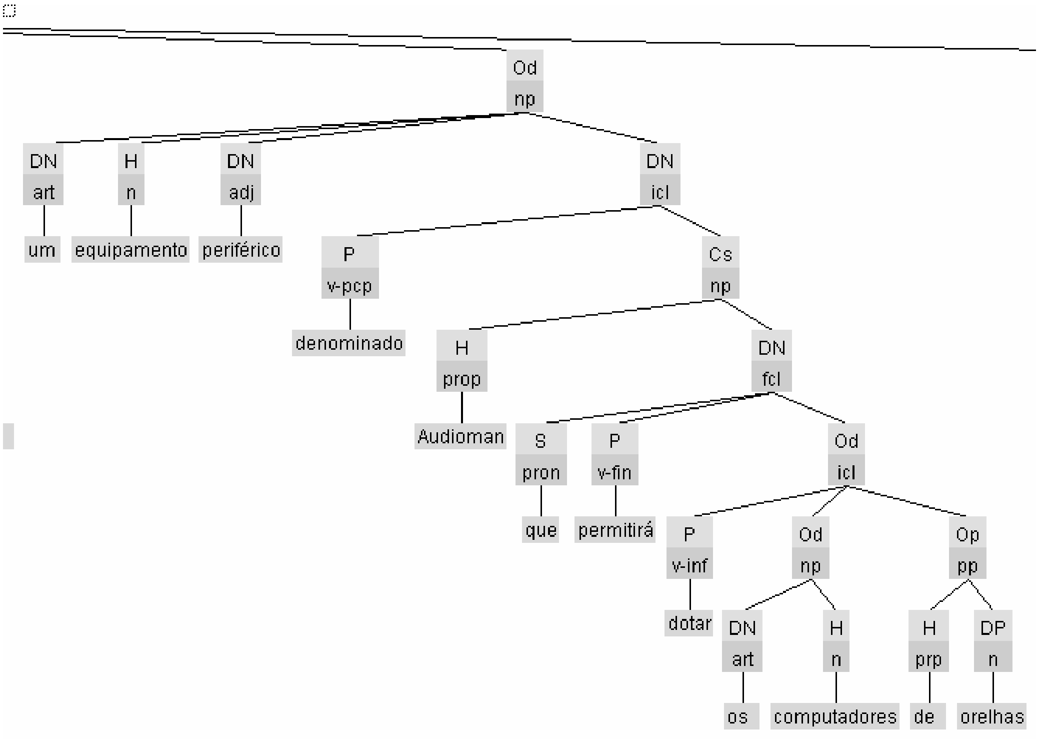
The second search example looks for clause hierarchies, here relative clauses with internal finite object clauses. Below, part of the resulting concordance and a corresponding tree section are shown:



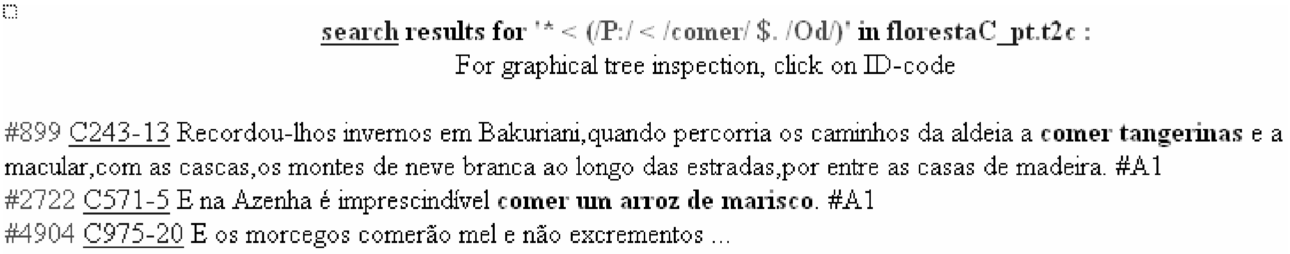
search results for '/DN:fcl/ < /Od:icl/' in florestaC_

For graphical tree inspection, click on ID-code

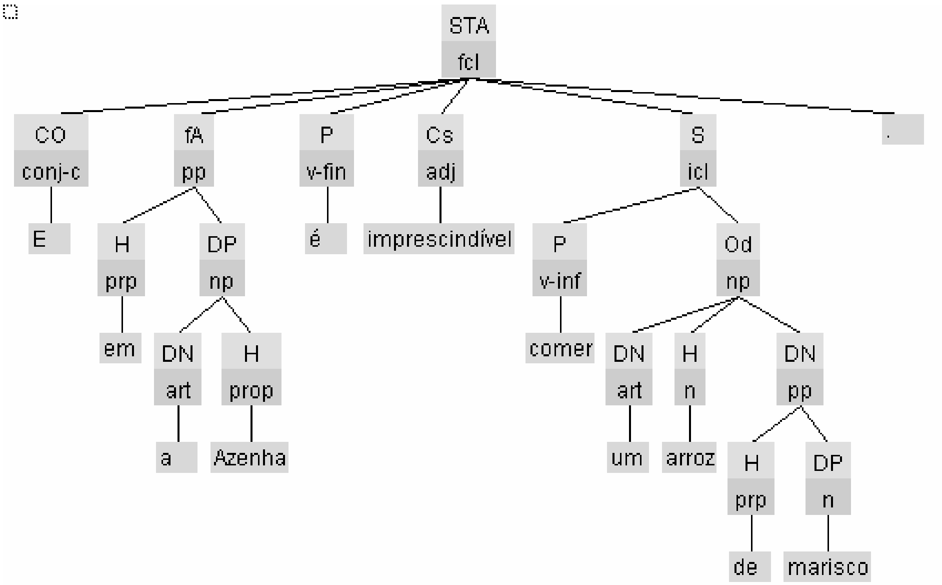
#122 C1-5 que querem fugir a algumas movimentações nocturnas já a caminho da ritualiza
ao Calypso e encontramo-nos na Locomia
#131 C3-2 que permitirá dotar os computadores de orelhas
#166 C9-1 que, há 3000 anos, decide abandonar a sua terra árida para se instalar no Egi
#205 C17-3 que esperam ver a situação clarificada, independentemente da parte que aca
#507 C185-3 em que a União Europeia decidiu abandonar a exploração do carvão de pe
#594 C202-5 que não conseguiram detectar a tempo o trabalho de sapa que Aldrich Am
#628 C207-5 que as máquinas de propaganda pretendem fazer crer
#663 C211-2 que nos permitam pensar que a frequência da depressão seja diferente no r
#798 C28-2 que permitem compatibilizar diversas redes de computadores, nomeadament
mesma empresa
#903 C243-17 que todos procuravam dar-lhe na Praça Vermelha no dia da vitória



The last example is "lexicographical", showing how to extract selection restrictions for objects of a given verb (here "comer").


search results for '* < (/P:/ < /comer/ $. /Od/)' in florestaC_pt.t2c :
For graphical tree inspection, click on ID-code

#899 C243-13 Recordou-lhos invernos em Bakuriani,quando percorria os caminhos da aldeia a **comer tangerinas** e a macular,com as cascas,os montes de neve branca ao longo das estradas,por entre as casas de madeira. #A1
#2722 C571-5 E na Azenha é imprescindível **comer um arroz de marisco**. #A1
#4904 C975-20 E os morcegos comerão mel e não excrementos ...



Acknowledgement

Floresta Sintá(c)tica is (partially) supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

References

Bick, Eckhard (2003-1). "A CG & PSG Hybrid Approach to Automatic Corpus Annotation". In: Kiril Simow & Petya Osenova (eds.), "Proceedings of SProLaC2003" (at Corpus Linguistics 2003, Lancaster), pp. 1-12

Christ, Oli (1994). "A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest.