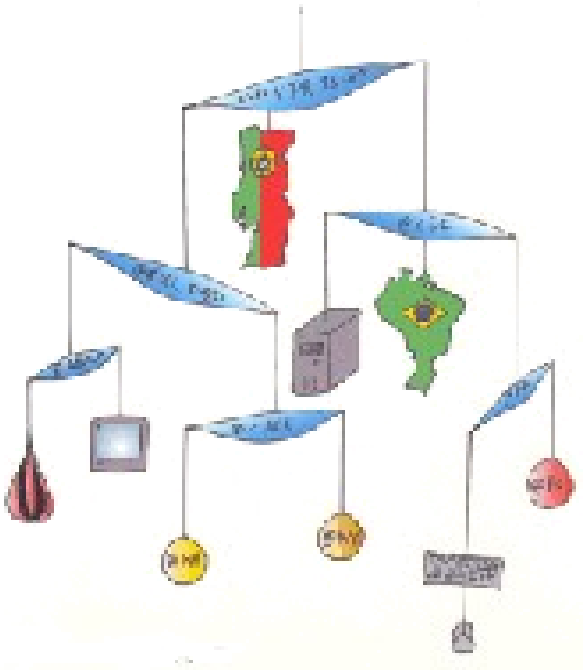
Automatic Syntactic Annotation

Eckhard Bick

University of Southern Denmark
Institute of Language and Communication
eckhard.bick@mail.dk

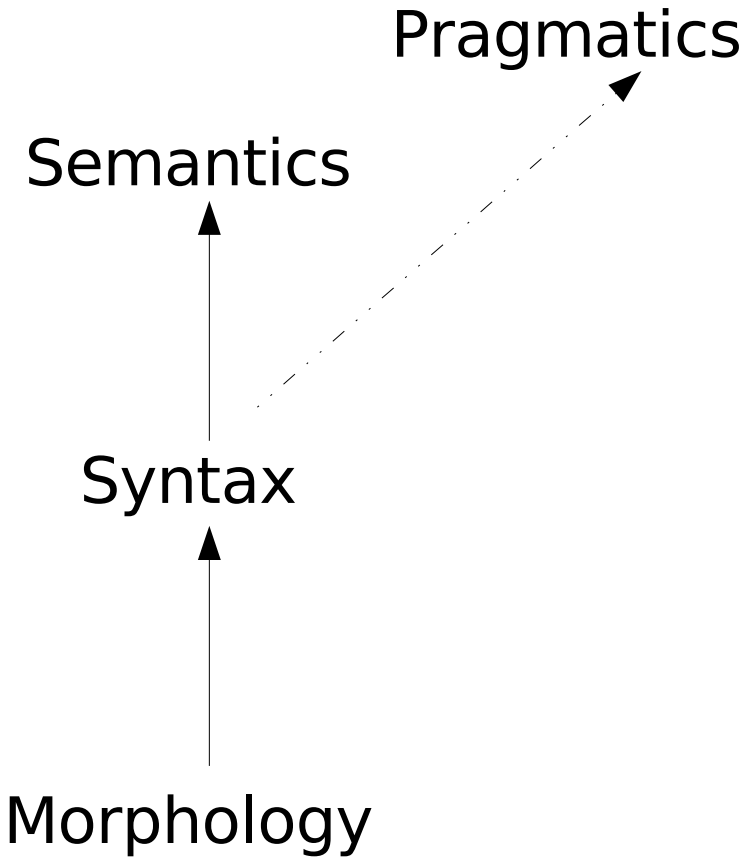


What is syntax?

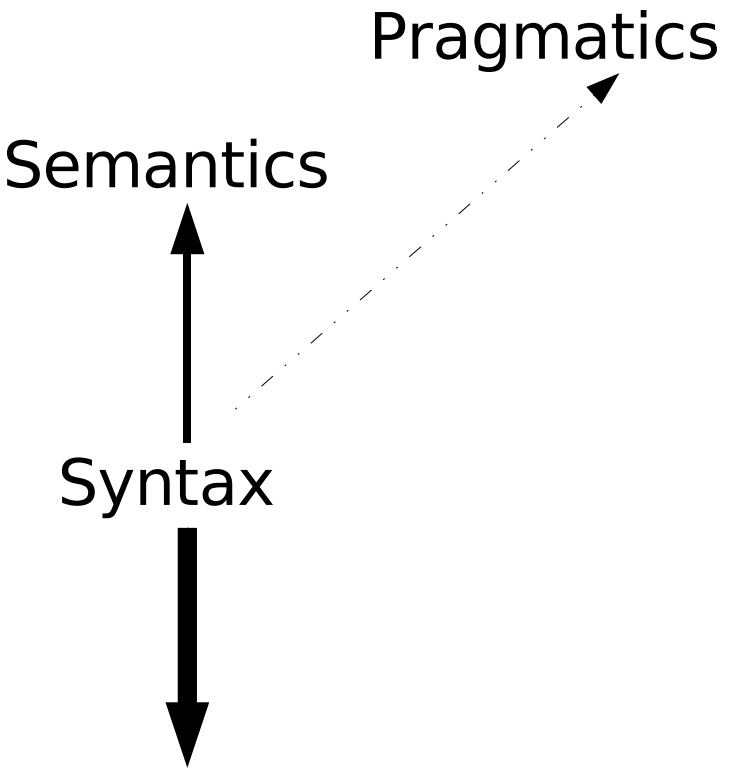
- combining words / tokens / morphemes ...
- rules for doing so (prescriptive vs. descriptive)
- syntactic form vs. syntactic function
  (clauses, phrases vs. subject, object etc.)
- a projection of semantics/pragmatics?
- an innate function of the Broca center?
- just another formalism for logical thinking?

Why syntax?

The onion layer hierarchy hypothesis:



Syntax as disambiguation pivot

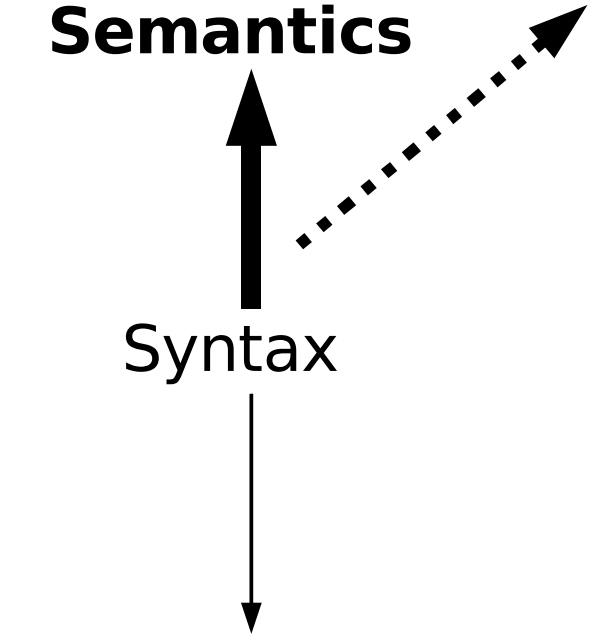


- conjunctions, relative pronouns
- participle or adjective: *chamado - querido - publicado*
- noun??: *os velhos/desilusionados/outros*
  o ficarmos aqui no Porto

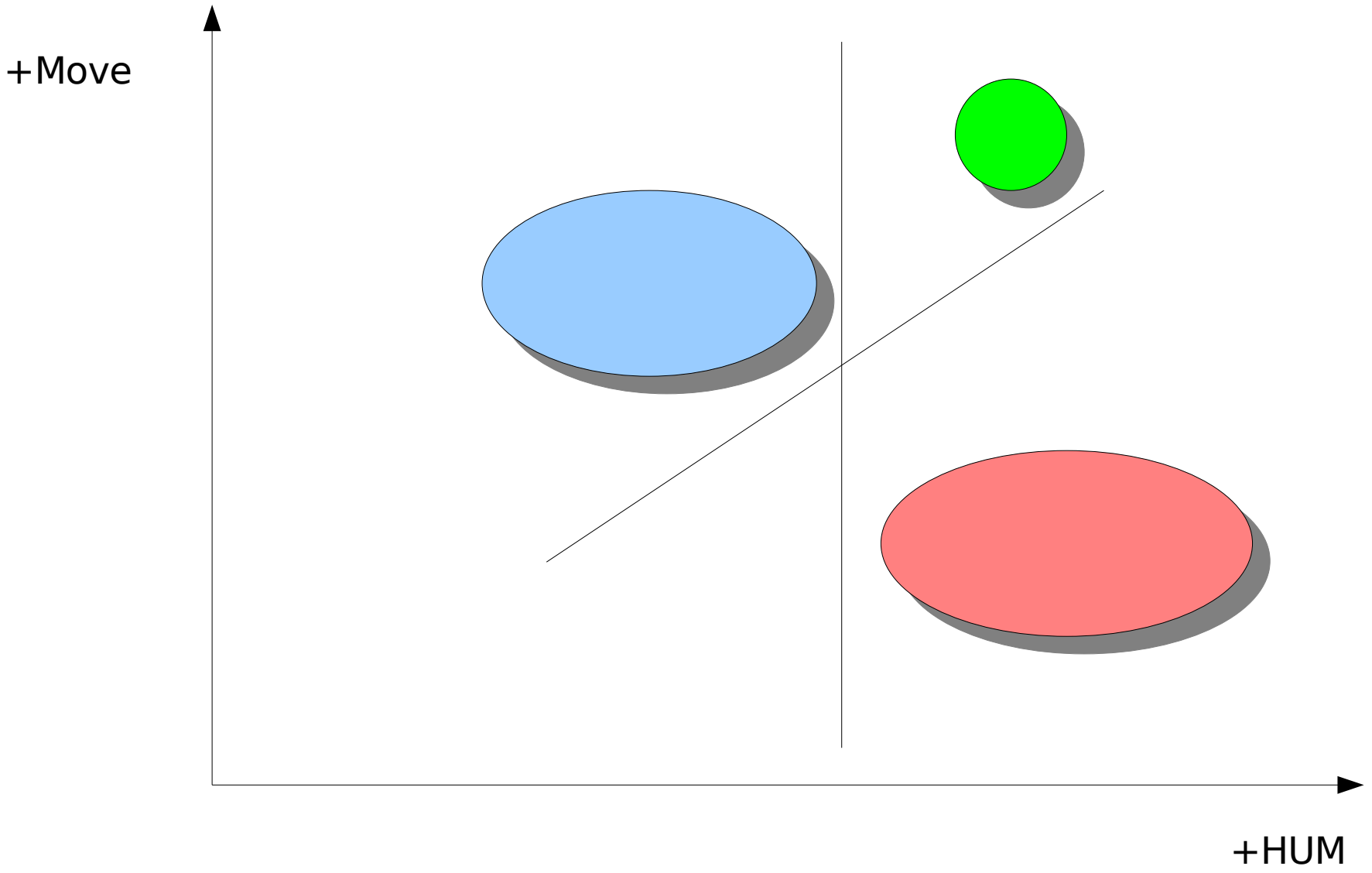
Syntax as disambiguation pivot

um homem grande - um grande homem
não adianta <vi> - adiantar a noção que <vt>
assente (based??) - assente em (based on)

Pragmatics:
word order to mark
question, request etc.



discrimination, not definition



Syntactic annotation styles

- Focus on syntactic form
  - Phrase structure grammar (PSG) -> labelled brackets
  - Dependency grammar (DG) -> labelled arcs
- Focus on syntactic function
  - Constraint grammar (CG) -> dependency pointers
- Focus on semantic function
  - Case roles (Filmore)
  - Lexical Functional Grammar (LFG)

Syntactic models

1. The flat classical model: word function, no form

O meu hipopótamo não come peixe.

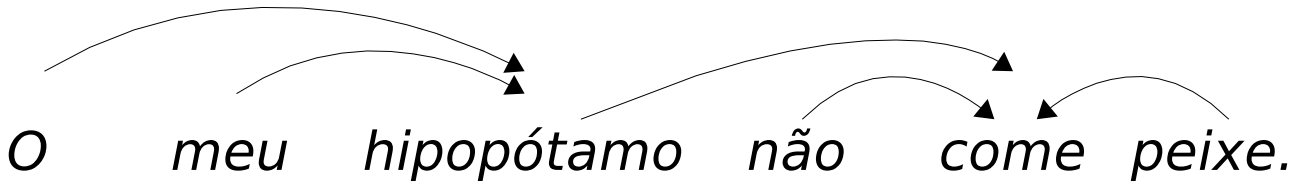
S A V O

- word-based
- psychologically easy to grasp
- function markers attached to semantically heavy words
- easy to turn into tags:

O	article	PRE-N
meu	determiner	PRE-N
hipopótamo	noun	S
não	adverb	A
come	verb	V
peixe	noun	O

2. Dependency grammar

- strictly token based (e.g. Tesnière)
  - expresses syntactic form as asymmetrical relations (“arcs”) between head tokens and dependent tokens
  - no zero tokens, no nonterminal nodes
- each dependent is allowed 1 head (exc. secondary arcs)
- directed acyclic graphs
- projective or non-projective (crossing branches / discontinuity)

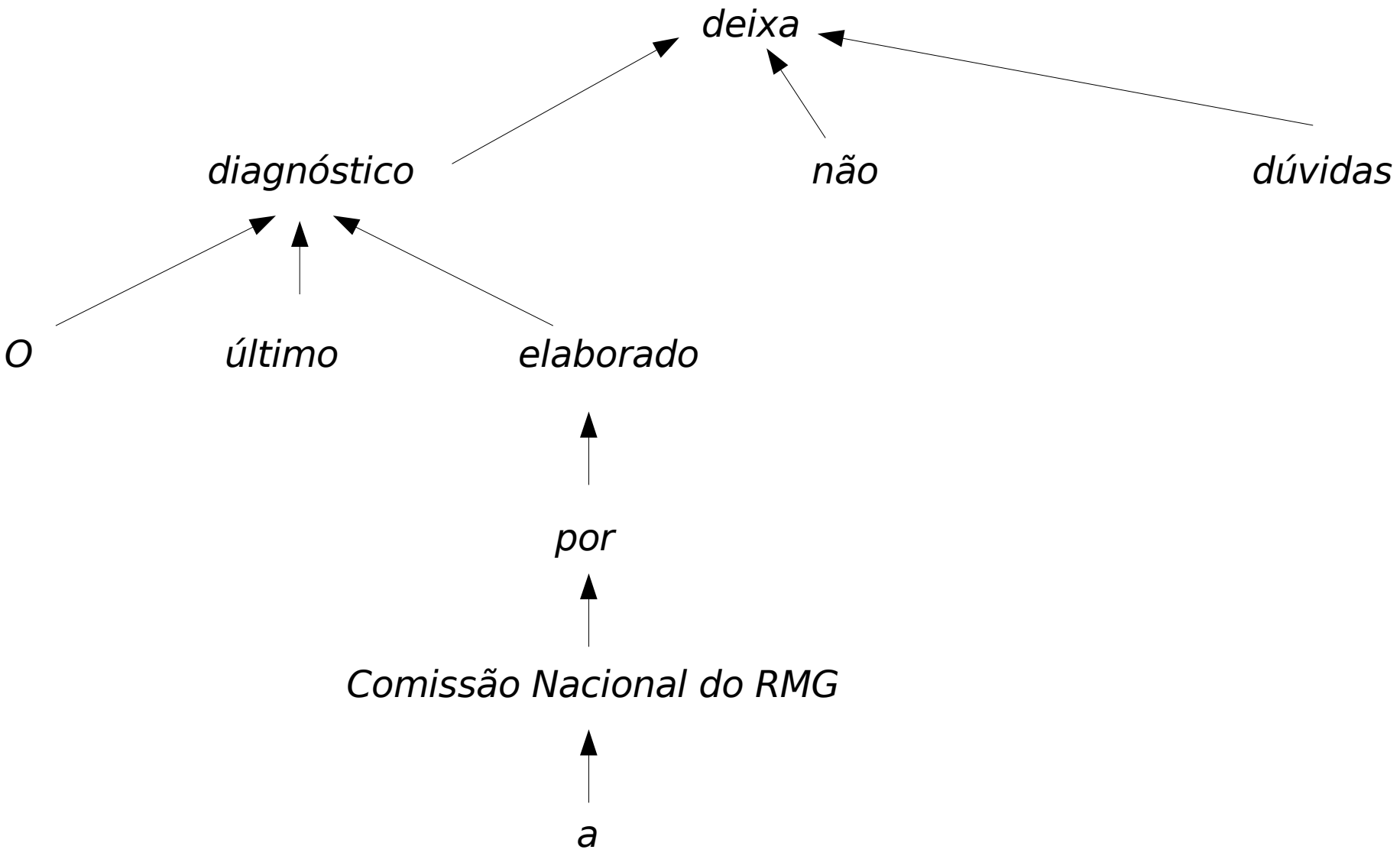


Dependency grammar annotation

O	#1->3
último	#2->3
diagnóstico	#3->9
elaborado	#4->3
por	#5->4
a	#6->7
Comissão=Nacional=do=RMG	#7->5
não	#8->9
deixa	#9->0 ROOT
dúvidas	#10->9

O	<id=1> <head=3>
último	<id=2> <head=3>
diagnóstico	<id=3> <head=9>
elaborado	<id=4> <head=3>
por	<id=5> <head=4>
a	<id=6> <head=7>
Comissão=Nacional=do=RMG	<id=8> <head=5>
não	<id=8> <head=9>
deixa	<id=9> <head=0> ROOT
dúvidas	<id=10><head=9>

Dependency grammar as trees



Dependency grammar with brackets “a la PSG”, e.g. TIGER

- **Penn-style:**

[V come [N hipopótamo [ART o][DET meu]] [A não] [N peixe]]

- **Vertical:**

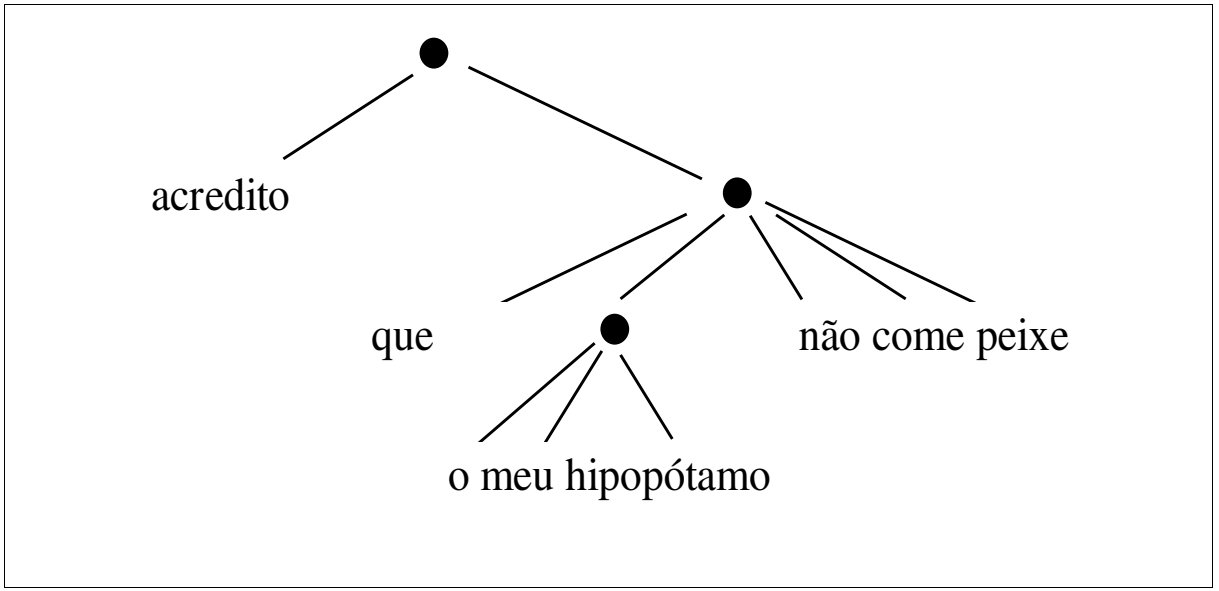
```
[V  come
      [N hipopótamo
            [ART o]
            [DET meu]
      ]
      [A não]
      [N peixe]
]
```

3. Constituent Grammar

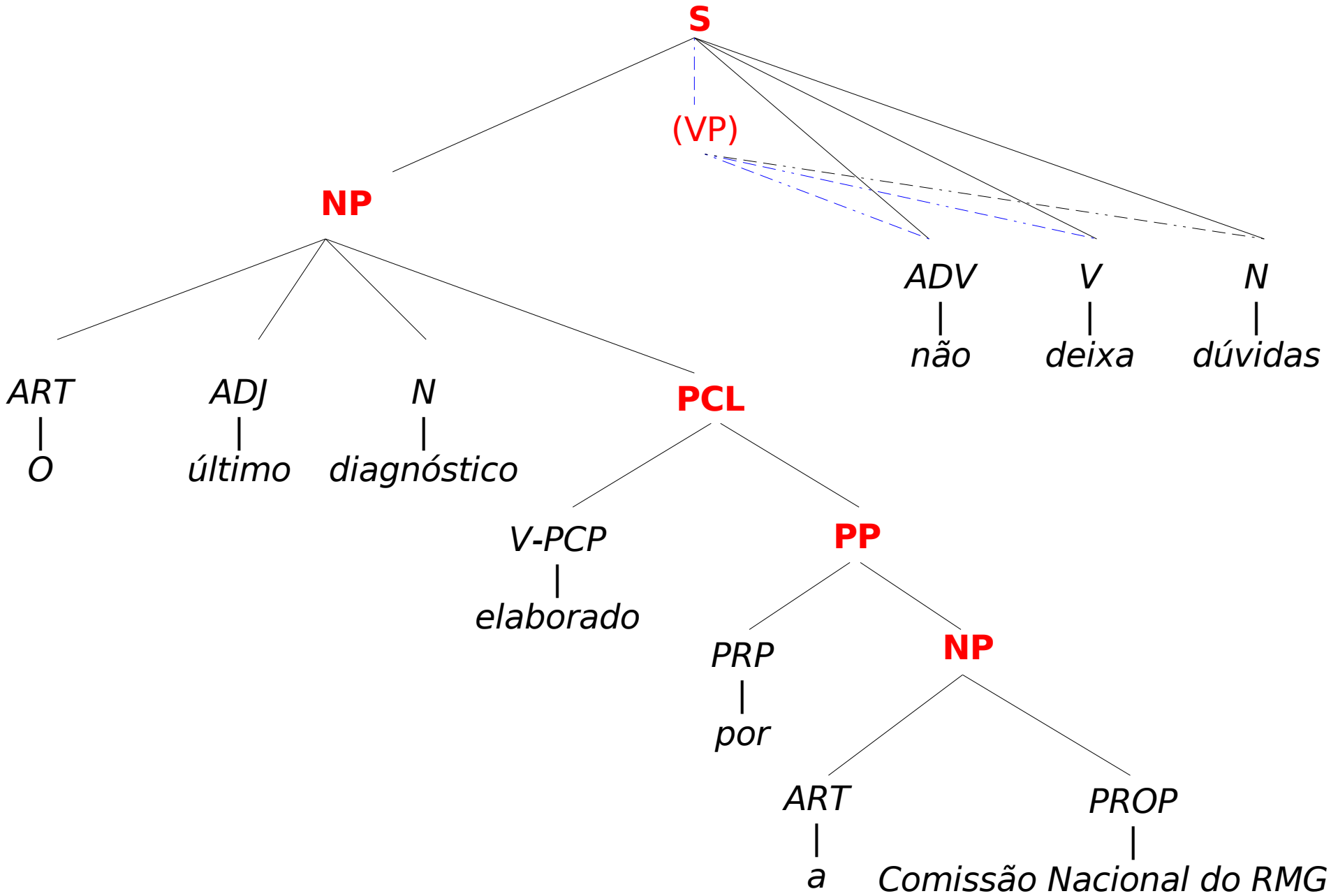
- hierarchical word grouping with non-terminals (e.g. Chomsky)
- syntactic form, no (or implicit) function
- expressed by rewriting rules, where a nont-terminal node is rewritten as a sequence of non-terminals and terminals (words or word classes)

s -> np vp
np -> art n
vp -> v np

Pure Constituent Grammar:



Classical PSG with phrase labels

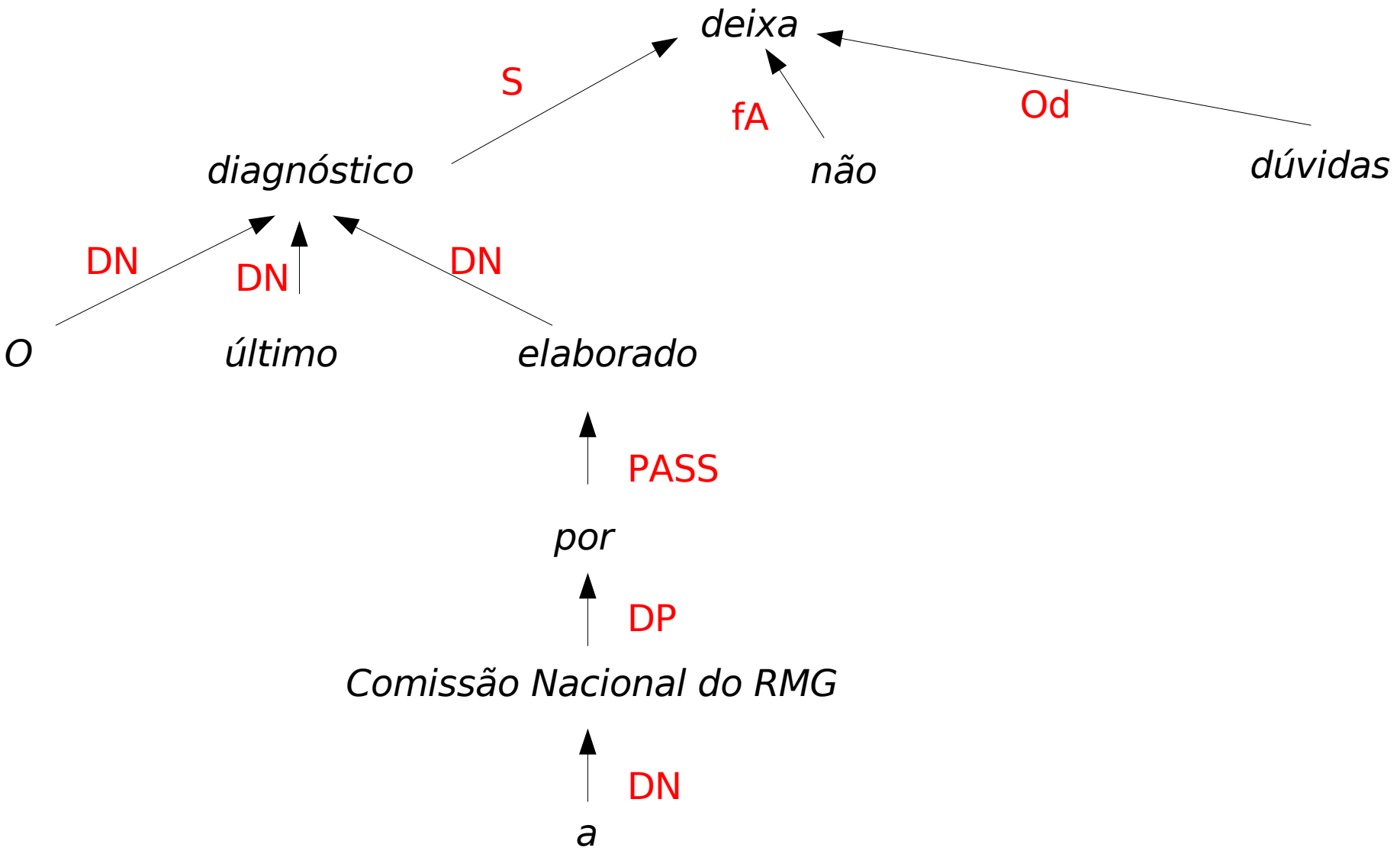


PSG annotation

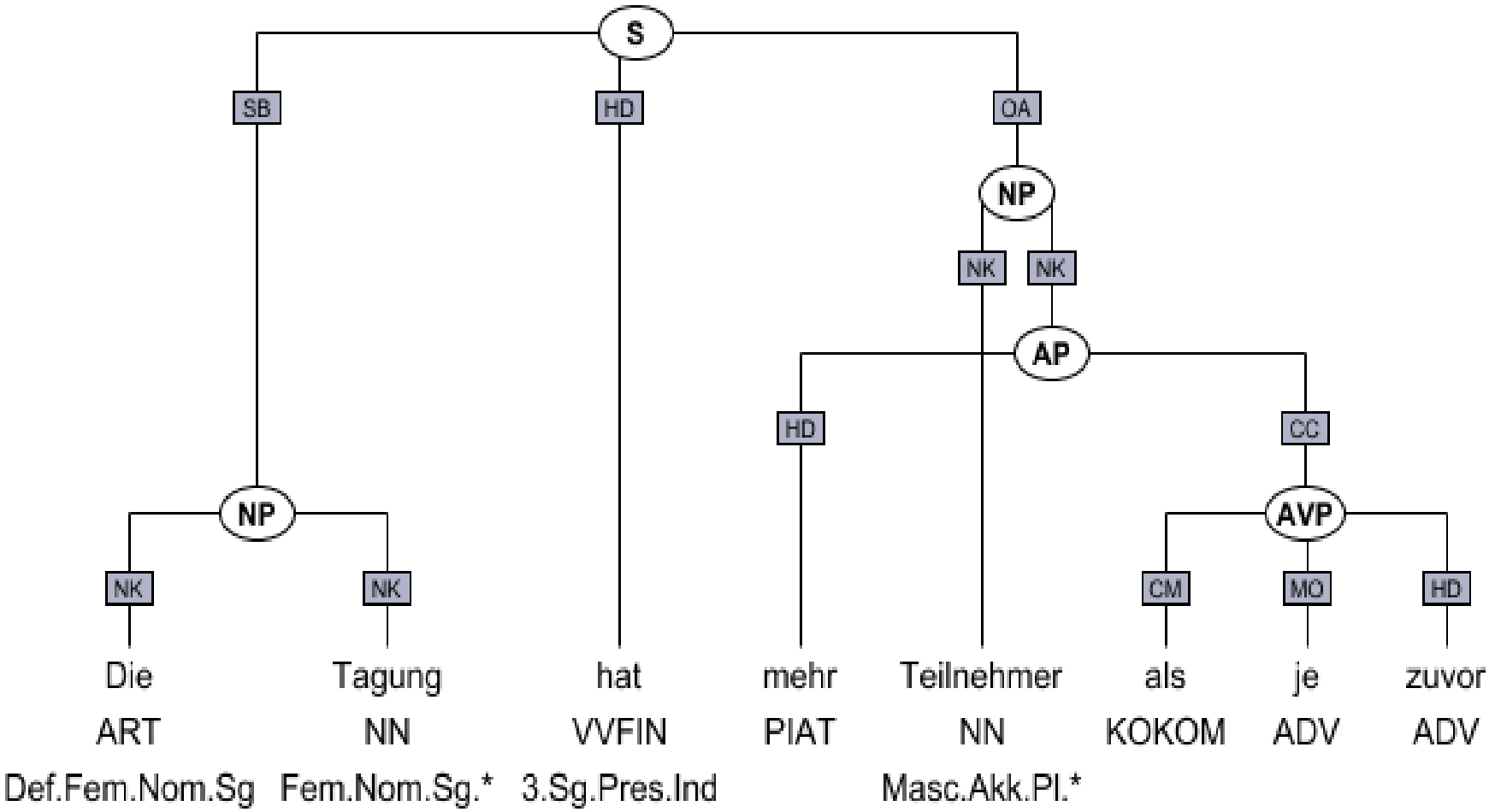
- **Penn Treebank bracketing**: Labeling opening brackets
  - [NP A minha irmã] [VP não fala [PP com [NP as amigas]]]
- **SUSANNE Treebank bracketing**: Labeling all brackets (cf. EAGLES)
  - [NP A minha irmã NP] [VP não fala [PP com [NP as amigas NP] PP] VP]
- **Vertical indented** (her with part of speech on one line):
  - [NP
    [Art A]
    [Det minha]
    [N irmã]
    NP]
    [VP
    [Adv não]
    [V fala]
    [PP
    [Prp com]
    [NP
    [Art as]
    [N amigas]
    NP]
    PP]
    VP]

Adding function:

- Dependency Grammar with function: adding function ("edge labelse") to dependency arcs



- Constituent Grammar with function:
  - NEGRA, TIGER: cat labels (mother) vs. edge label (daughter

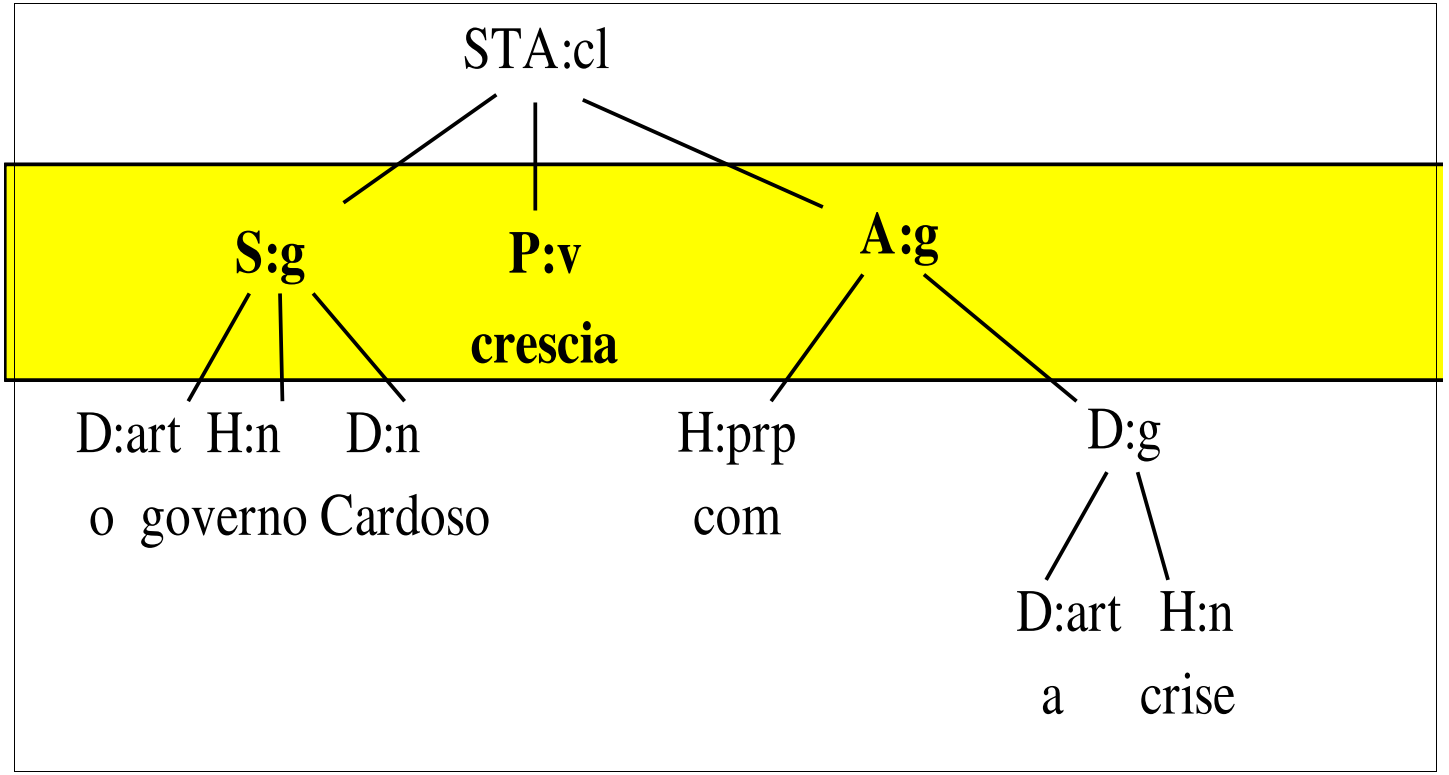


- Constituent Grammar with function:
  - VISL (function:form labels for each node)

Vertical Notation:

```
STA:cl
S:np
=DN:art     O
=H:n        governo
=DN:prop    Cardoso
P:v-fin
A:pp
=H:prp      com
=DP:np
==DN:art    a
==H:n       crise
```

Graphical Notation:



PSG (Chomsky)	**DG** (Tesnière, Melcuk)
Definite Clause Grammar (DCG) - Prolog *Transformational-Generative Grammar (TGG)* *Head-Driven Phrase Structure Grammar (HPSG) Functional Unification Grammar (FUG),* *Lexical Functional Grammar (LFG)* *Tree Adjoining Grammar (TAG) - Arvid Joshi* *AGFL* ... *Functional Grammar (FG) - Simon Dik* *Systemic Functional Grammar (SFG) - Halliday*	*Constraint Grammar (CG)* *Functional Dependency Grammar (FDG)* *Topological Dependency Grammar (TDG)* *Extensible Dependency Grammar (XDG)* *Link Grammar (D-H symmetry)* *Dependency Grammar Annotator (DGA)*
constituent based structure	token based structure
explanatory-linguistic perspective generative tradition: parsers	descriptive-applicational perspective analytical tradition: taggers
(labeled) brackets	(labeled) arcs
rewriting rules	attachment rules
originally fixed word order languages • English, French	originally free word order languages • Slavonic languages, Finnish, German
Problems: Discontinuity, free word order	Problems: Coordination, ellipsis
declarative programming	procedural programming
linguist-written: AGFL, HPSG, VISL-PSG, PATR, XTAG, ... machine-learned: PCFG e.g. Viterbi, Collins, Bikel	linguist-written: ENGCG, GERCG (Lingsoft), Machinese parsers (Conexor), VISL parsers machine-learned: MALT, Matsumoto, MSTParser,

	Constituent treebanks	Dependency treebanks
English	Penn I & II, Susanne Corpus, TOSCA Lancater/IBM treebank (Spoken E.)	English Dependency Treebank and The PARC 700 Dependency Bank and CHILDES (Brown)
Arabic	Penn Arabic Treebank	PDT-Arabic (Smrž et.al. 2002)
Basque		Eus3LB
Bulgarian	BulTreeBank(Simov et.al. 2005) HPSG	-->
Catalan		Cat3LB
Chinese	Penn Chinese Treebank	Sinica Treebank(Chen et.al. 2003)
Czech		PDT (Hajič et.al. 2001)
Danish	Arboretum (Bick 2003)	Arboretum (Bick 2003), Danish Dependency Treebank (Kromann 2003)
Dutch	Corpus Gresproken Nederlands	Alpino Treebank (van der Beek et.al. 2002)
Estonian	Arborest	
French	IBM Paris Treebank, Abeillé Treebank, L'Arboratoire	
German	NEGRA, TIGER Treebank (Brants et.al 2002), TueBa-D/S (spoken, +topology), TueBa-D/Z (written)	-->
Greek		Greek Dependency Treebank
Hebrew	Hebrew Treebank (a la Penn)	
Hungarian	(project: Hungarian treebank)	
Italian	VIT - Venice Italian Treebank	Turin University Treebank (Bosco et.al.)
Japanese	VERBMOBIL (Kawata and Bartels 2000)	ATR Dependency corpus (Lepage et.al.)
Korean	(project: Korean Treebank)	
Norwegian	(project: TREPIL Norwegian treebank)	
Polish	(HPSG test suite treebank)	
Portuguese	Floresta Sintá(c)tica (Afonso et.al. 2002)	-->
Slovene		Slovene Dependency Treebank (Džerosky et.al. 2006)

4. Constraint Grammar (CG)

- CG as a **descriptive paradigm**
  - function-first approach with token-based function tags
  - Classic CG: shallow depedency (attachment direction, head type)
  - depth and constituents only implicitly marked

O	@>N	(pointer to head type: N)
meu	@>N	
hipopótamo	@SUBJ>	(direction pointer without head type)
não	@ADVL>	
come	@FMV	(top node)
peixe	@<ACC	

Adding full **numbered dependency**

- Integrated formalism: FDG
- Add-on attachment rules: PALAVRAS

@<ACC --> (<mv>) IF (L)

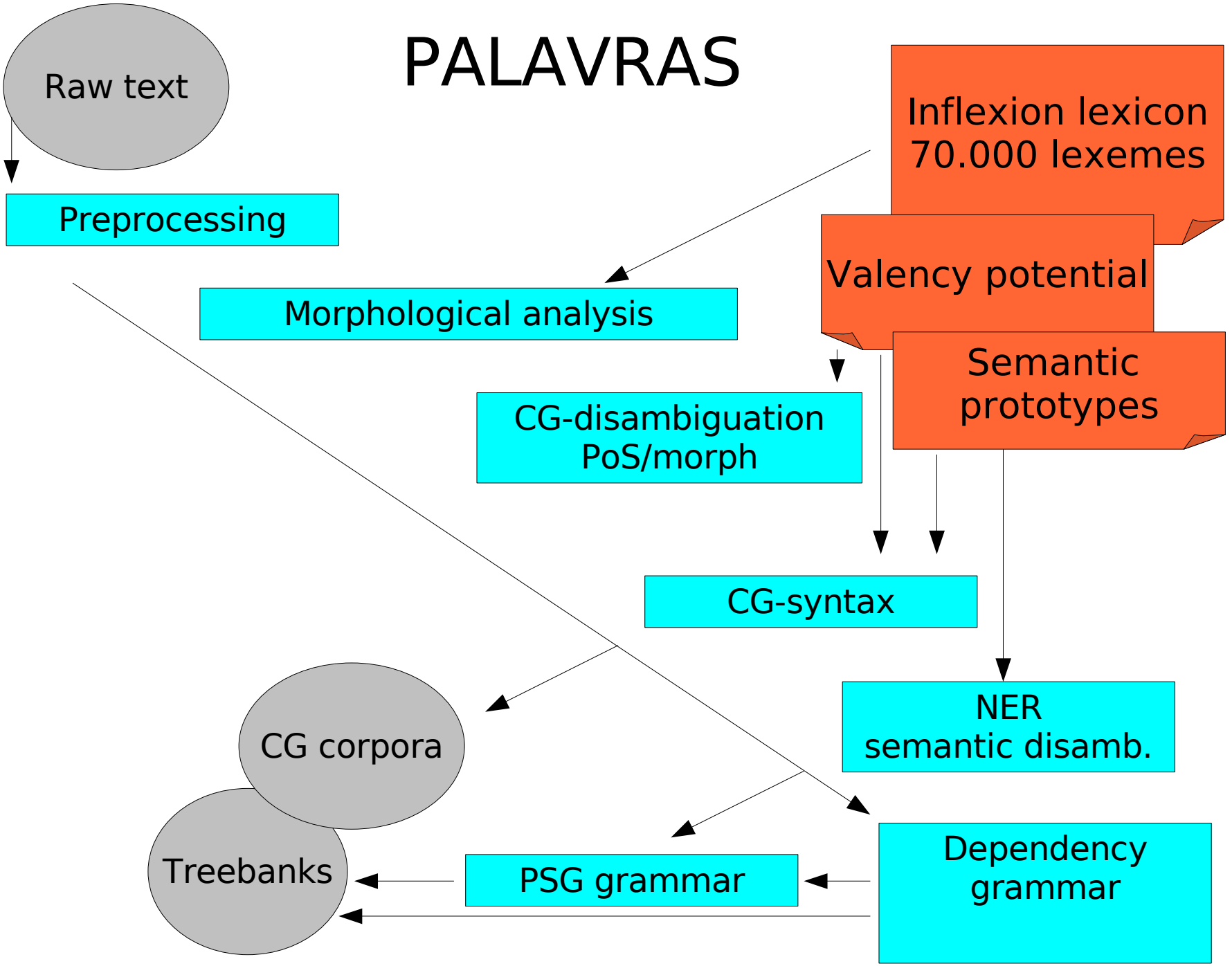
@SUBJ> --> (VFIN) IF (R) BARRIER:(@FS)

<np-long> --> (N,PROP,PERS,INDP,¤NP-HEAD)
IF (L) HEADCHILD=(<np-close>)

O <artd> DET M S	@>N	#1->3
último ADJ M S	@>N	#2->3
diagnóstico N M S	@SUBJ>	#3->9
elaborado V PCP2 M S	@ICL-N<	#4->3
por PRP	@<PASS	#5->4
a <artd> DET F S	@>N	#6->7
Comissão=Nacional PROP F S	@P<	#7->5
não ADV	@ADVL>	#8->9
deixa V PR 3S	@FMV	#9->0
dúvidas N F P	@<ACC	#10->9
$.		#11->0

Taggers and parsers for Portuguese

- **PALAVRAS**: CG-based robust DG & PSG parser http://visl.hum.sdu.dk/itwebsite/port/portgram.html (Bick 2000)
- **Curupira**: Robust syntactic parser, based on ranked and constrained ReGra PSG rewriting rules http://www.nilc.icmc.usp.br/nilc/tools/curupira.html (Martins, Hasegawa & Nunes 2002)
- QTAG-based PoS-tagger for Brazilian Portuguese, trained on 500M words, Precision =93% http://lael.pucsp.br/corpora/etiquetagem/ (Sardinha & Lima-Lopes)
- FreP - Phonlological analysis at the word level and below http://www.fl.ul.pt/LaboratorioFonetica/frep/
- **GojolParser** http://www.linguateca.pt/Repositorio/GojolParser.txt, DG & PSG, commercial, calls itself the best (error rate < 1%)
- Hermes - Tokenizer and PoS tagger http://hermes.sourceforge.net/hermesweb.html (FURG, open source)
- jspell - morphological analyzer http://natura.di.uminho.pt/natura/natura?&topic=jspell (Projecto Natura - U.Minho, Linguateca ...)
- LX-Suite - lemmatizer and PoS tagger, parser (LX-Gram planned for syntax - http://lxsuite.di.fc.ul.pt/ (NLX group, University of Lisbon)
- PoSiTagger - symbolic PoS tagger http://www.nilc.icmc.usp.br/nilc/projects/mestradorachel.html (Aires & Aluísio 2000)
- TreeTagger - a language independent PoS tagger (Schmid & Stein) http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html, trained for Portuguese http://gramatica.usc.es/~gamallo/tagger.htm (Pablo Gamallo)
- Xerox PoS tagger - twol with HMM-disambiguation http://www.xrce.xerox.com/competencies/content-analysis/demos/portuguese


PALAVRAS
Raw text
Preprocessing
Inflexion lexicon
70.000 lexemes
Valency potential
Semantic
prototypes
Morphological analysis
CG-disambiguation
PoS/morph
CG-syntax
NER
semantic disamb.
CG corpora
Treebanks
PSG grammar
Dependency
grammar

- CG as a **methodological paradigm**
  - reductionist: focus on disambiguation, constraints as to what is *not* allowed in a given context
  - progressive level annotation: same method and tag-based annotation for ever higher linguistic levels
    - lexicon
    - morphology (“Analyzer”, “multitagger”, cohorts)
    - PoS disambiguation (“tagger”)
    - syntactic potential/mapping
    - syntactic disambiguation (“parser”, PALAVRAS syntax)
    - precise attachment (dependency or constituent structure)
    - case roles, clause bounderies, semantic classes, valency instantiation, anaphora resolution, discourse markers .... --> add your own module!
  - services many different NLP applications: Corpus research, MT, teaching, spellchecking, QA ...

A Constraint Grammar rules file

DELIMITERS (1 line, defines sentence boundaries)

DELIMITERS = “<.>” “<!>” "<?>" ;

SETS (1 or more sections of set definitions)

LIST N-LOC = <inst> <L> <Lh> <Lciv> <Lwater> <Lpath> <build> <BB> ;

LIST PROP-LOC = <top> <civ> <inst>

SET N/PROP-LOC = N-LOC OR PROP-LOC

MAPPINGS (adding new tags, e.g. syntax)

MAP (@SUBJ>) TARGET N/PROP (*-1 >>> BARRIER NON-PRE-N) (1C VFIN)

MAP (%TOP-PL) TARGET ("em") IF (0 @ADVL) (*1 @P< LINK 0 N/PROP-LOC) ;

CORRECTIONS (replacing tags anywhere in a reading)

SUBSTITUTE (TAG-1) (TAG-2) TARGET (TAG-3) IF (Context1) .. (Context2)

CONSTRAINTS (1 or more sections of REMOVE or SELECT rules, with each section compiled and run seperately)

REMOVE (VFIN) (*-1C CLB-WORD) (*1C VFIN BARRIER CLB OR KC)

SELECT (N) (-1 (<artd>)) (1 (<KOMP>)) (2 (ADJ) OR (PCP)) ;

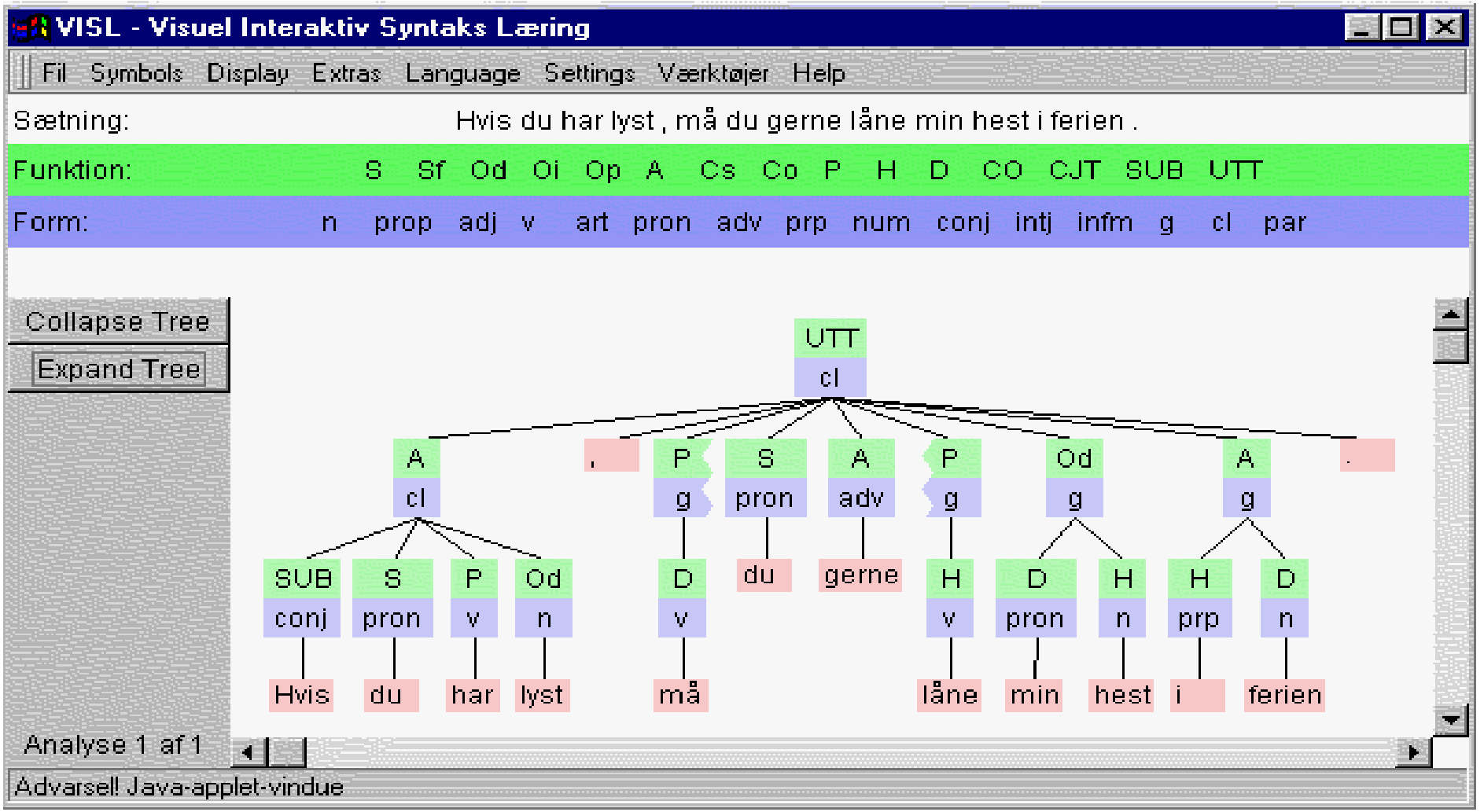
Applications for a syntactic parser

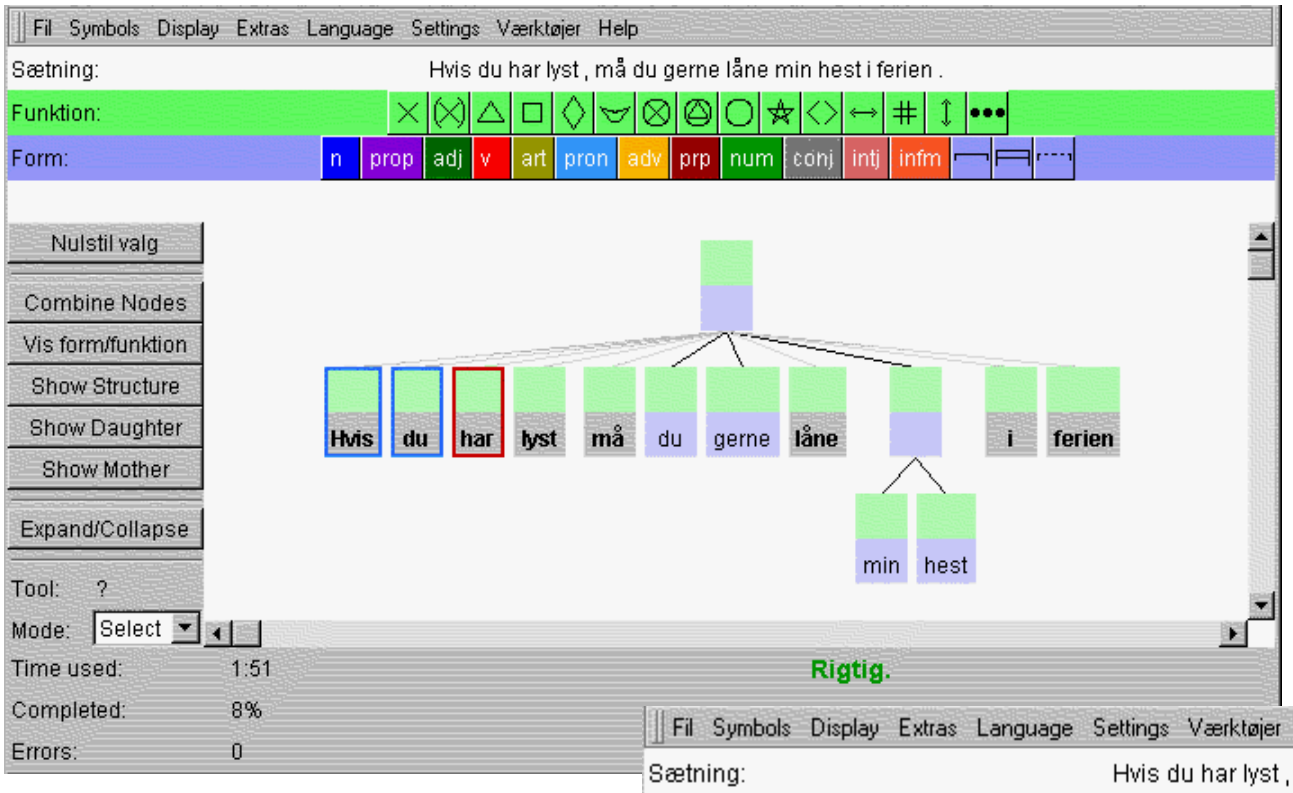
- Corpus annotation (Wednesday)
  - Linguistic research: Examples, statistics, comparative
  - Teaching: Empirical approach, language awareness
  - Revised data (Treebanks): machine learning
- NER (tomorrow)
  - Name chain recognition, e.g. PP @N< in institution names
  - semantic type inheritance from nouns (@APP, @SC)
  - Semantic type projection from valency slots (e.g. +HUM subject condition, +HUM attributes)

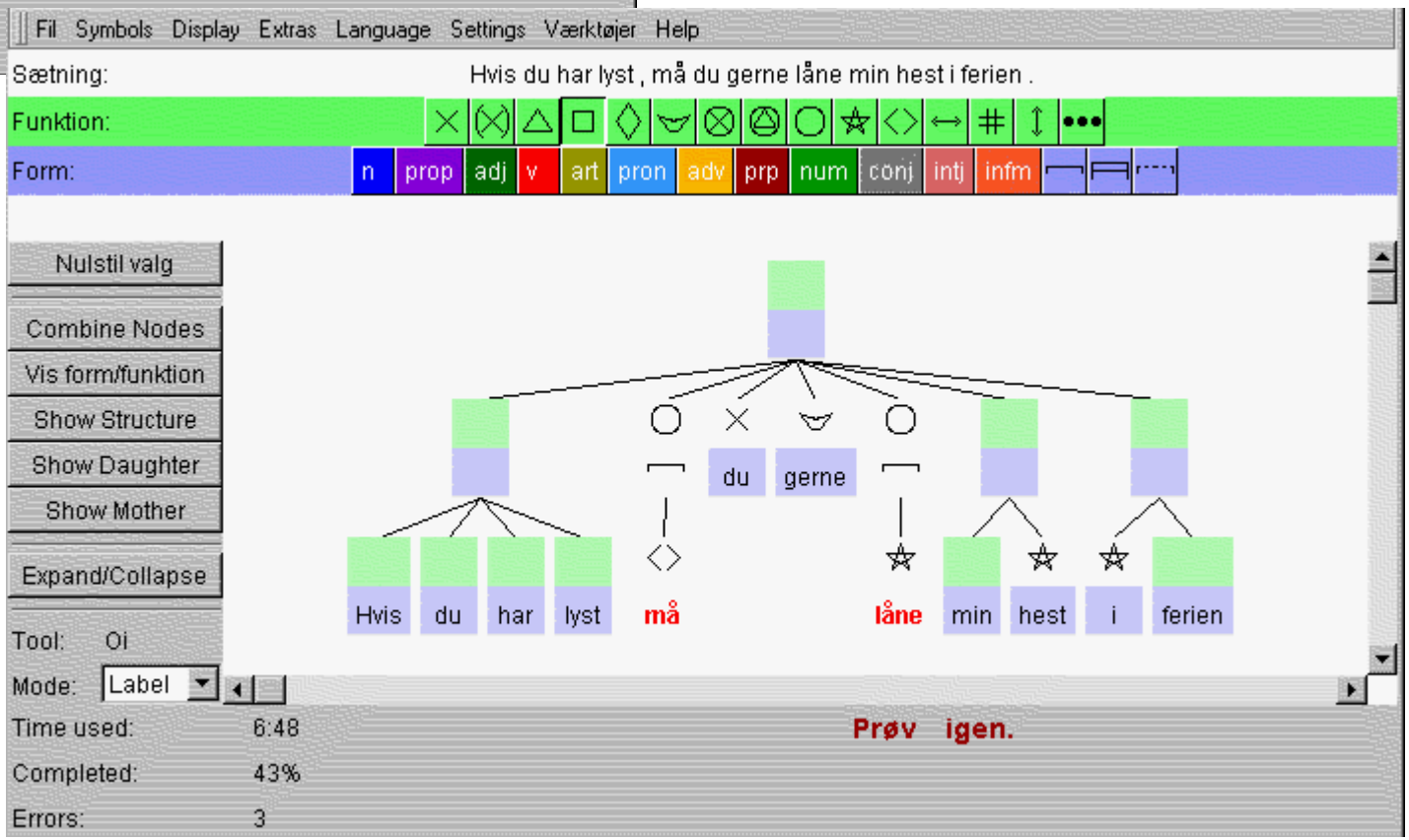
Teaching: e.g. VISL tools

1. TextPainter

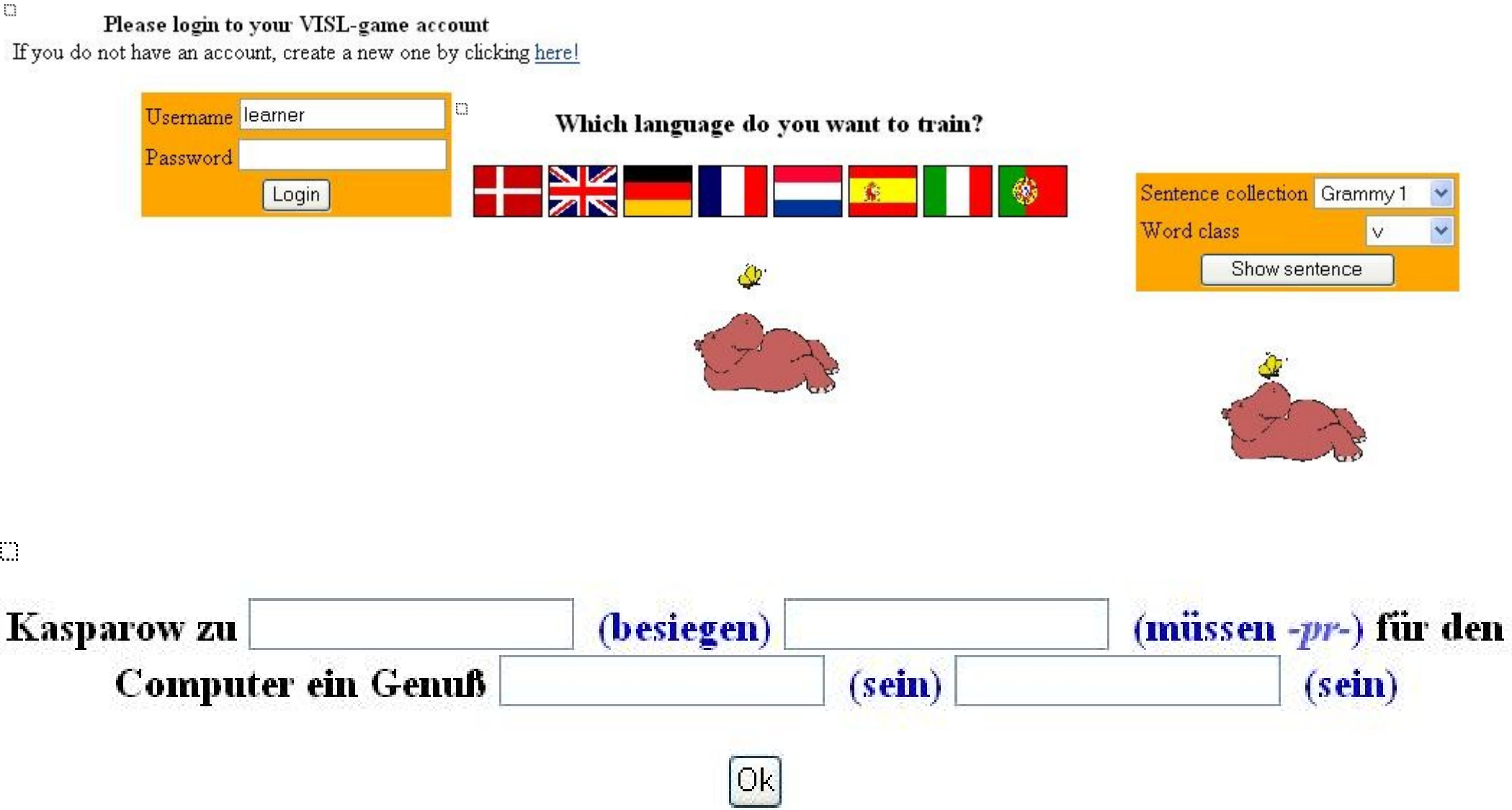
2. Interactive syntactic tree building




Fil Symbols Display Extras Language Settings Værktøjer Help
Sætning: Hvis du har lyst , må du gerne låne min hest i ferien .
Funktion:
Form: n prop adj v art pron adv prp num conj intj infm
Nulstil valg
Combine Nodes
Vis form/funktion
Show Structure
Show Daughter
Show Mother
Expand/Collapse
Hvis du har lyst må du gerne låne min hest i ferien
Tool: ?
Mode: Select
Time used: 1:51
Completed: 8%
Errors: 0
Rigtig.


Fil Symbols Display Extras Language Settings Værktøjer Help
Sætning: Hvis du har lyst , må du gerne låne min hest i ferien .
Funktion:
Form: n prop adj v art pron adv prp num conj intj infm
Nulstil valg
Combine Nodes
Vis form/funktion
Show Structure
Show Daughter
Show Mother
Expand/Collapse
Hvis du har lyst må du gerne låne min hest i ferien
Tool: Oi
Mode: Label
Time used: 6:48
Completed: 43%
Errors: 3
Prøv igen.

3. KillerFiller:
Automatic corpus-based slot-filler exercises



Question-answering systems (EPIA2003): better question-typing

QUE:fcl

=ADVL:adv('quando' <interr>) **Quando**

[=FOC:adv('é=que') **foi=que**]

=P:v-fin('nascer' PS 3S IND) **nasceu**

=SUBJ:prop('Balladur' <hum> M/F S) **Balladur**

=?

From this information the system fills in a number of variables:

question pattern (Atemp-PS)

interrogative constituent: Q-word ("quando"), Q-function ("ADVL")

predicator information: P-base ("nascer"), P-tense ("PS")

search point constituent: S-string ("Balladur"), S-function ("SUBJ"), S-head ("Balladur")

Hit sentence: *Balladur nasceu em Esmirna (Turquia), em 1929, e formou-se na Escola Nacional de Administração, de onde saiu a elite da função pública francesa.*

STA:cu
CJT:fcl
=**SUBJ:prop**('Balladur' <hum> M/F S) Balladur
=**P:v-fin**('nascer' PS 3S IND) nasceu
=ADVL:pp
==H:prp('em') em
==P<:np
===H:prop('Esmirna' <civ> M S) Esmirna
===(
====**N<PRED:prop('Turquia' <civ> F S) Turquia**
===)
=,
=**ADVL:pp**
==H:prp('em') em
==P<:num('1929' <date> <card> M S) 1929
=,

syntactic analysis permits to extract more implicit knowledge, e.g. ISA relations from appositions, predicatives and relative clauses:

1. Onde é/fica Smirna

2. Quando Rakhmonov derrubou o governo?

*A guerra civil no Tadjiquistão, que fez mais de 50 mortos, começou em 1992, quando as **forças** do neo-comunista Rakhmonov derrubaram o governo dos islamistas ...*

```
SUBJ:np
=H:n(<HH>)   forças
=N<:pp
==H:prp   de
==P<:np
===>N:art     o
===H:n(<hum>)  neo-comunista
===N<:prop(<hum>) Rakhmonov
```

- (a) name-np-flattening: post-nominal or appositive names are substituted for the np, whose head they are dependent of: O neo-comunista Rakhmonov -> Rakhmonov
- (b) toto-pro-pars: semantic heads of postnominal de-pp's are substituted for the pp: as forças de Rakhmonov -> as forças Rakhmonov

Apply a - b - a

Spell Checking: e.g. *OrdRet*

- Beyond list checking: marking “real” words as wrong in-context, e.g. infinitive vs. finite verb, number or gender agreement errors between subject, verb and/or predicative
  *han kunne ikke **finder** nøglen -> **finde***
  (he could not finds the key)
- weighting correction suggestions (very important for dyslexics and other bad readers): syntactically/contextually ones first
  *det **rejner** (it rains)*
  -> 1. regner (V PR: rains)
  2. rejer (N P: shrimps)
  3. rener (N P), 4. regne (INF), 5. rene (ADJ) ...
- marking structural errors, e.g. SV inversion, object order
  *i går **det regnede** en del (yesterday it rained a lot)*
  *-> i går **regnede det** en del*

Machine Translation: Polysemy resolution, Lexical transfer

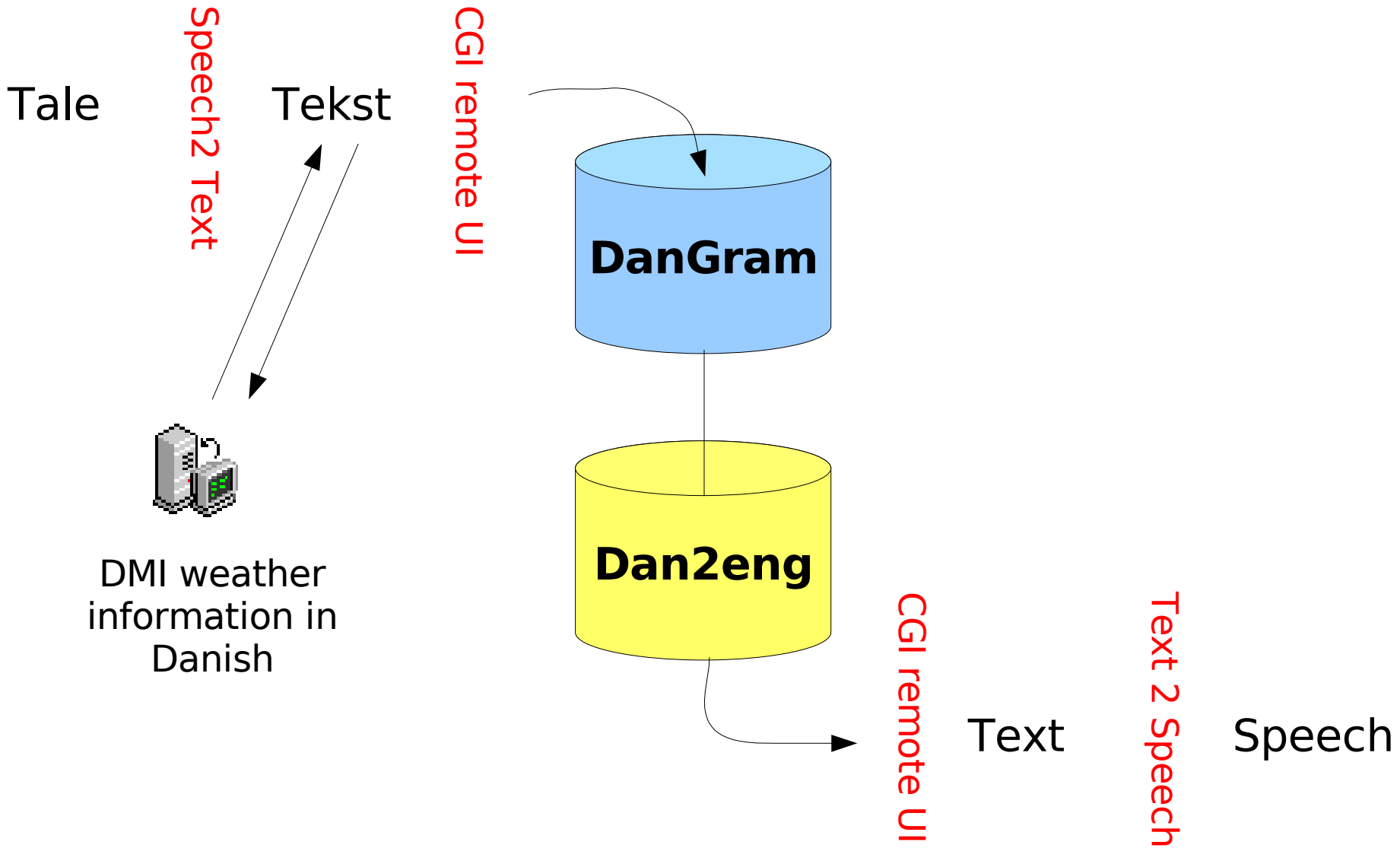
udsætte_V

- {opsætte} :postpone, :put=off;
- D=(@ACC) D=("for")_to :expose;
- D=(<prize> @ACC) :offer;
- S=(INF) M=(<quant>) :criticize;
- D=("vagt")_sentry :post;
- D=(<Vwater> @ACC) :put=out;
- D=("lejer" @ACC) :evict;

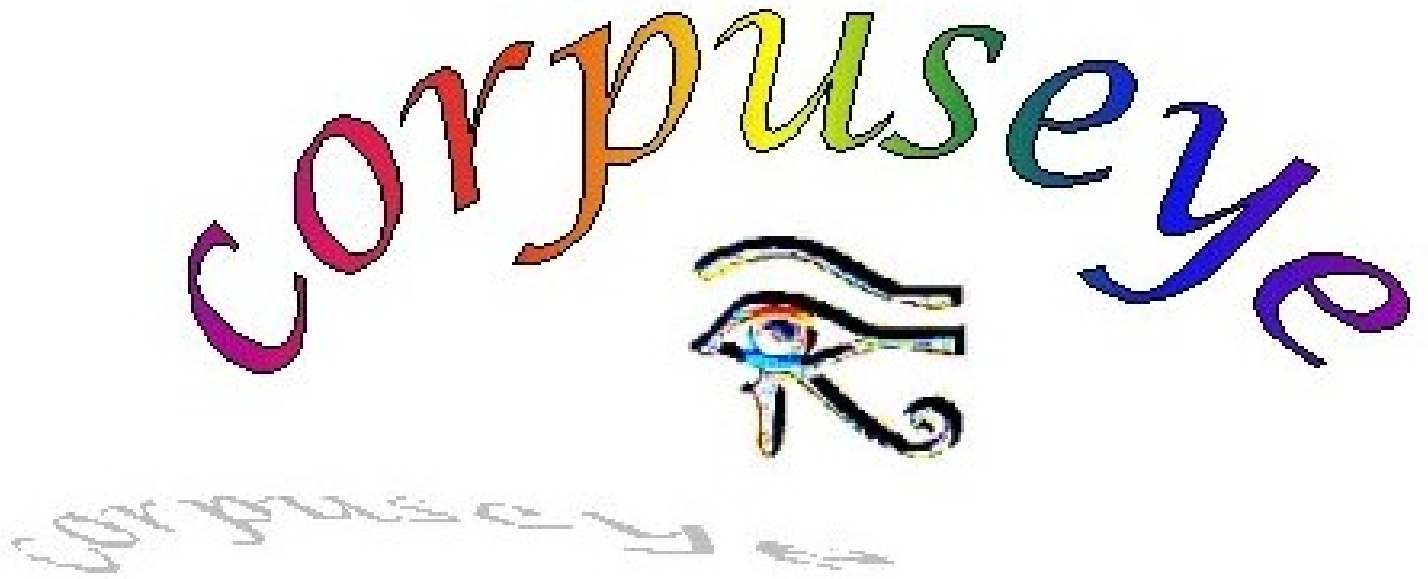
Machine Translation: Movement rules, Structural transfer

- ***I dag*** *@**ADVL*** *drikker @FMV vi @SUBJ vin @ACC -*
  Today *we drink wine*
- *(@ADVL|@ACC|@FS-ADVL|@FS-ACC|@>>P), I_dag*
  w(@FMV|@FAUX|@FS-[^Q]+), drikker
  w(@ICL-AUX<)?,
  w(@ADVL)?,
  (@SUBJ|@F-SUBJ|@S-SUBJ) vi
  -> 1, 5, 2, 3, 4

How to use the MT system: e.g. Weather forecasts for tourists



A user-friendly Corpus interface



standard search interface (old)

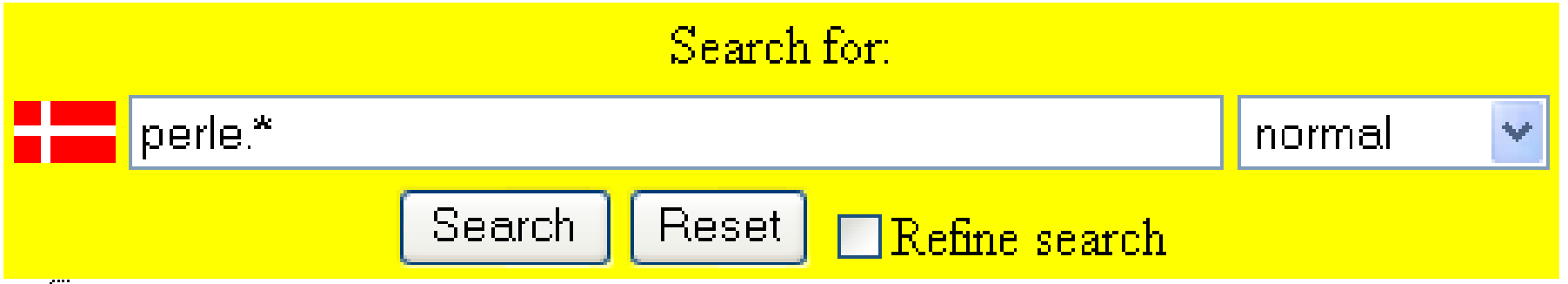
user-friendly cqp (new)

Treebanks

Guided tour

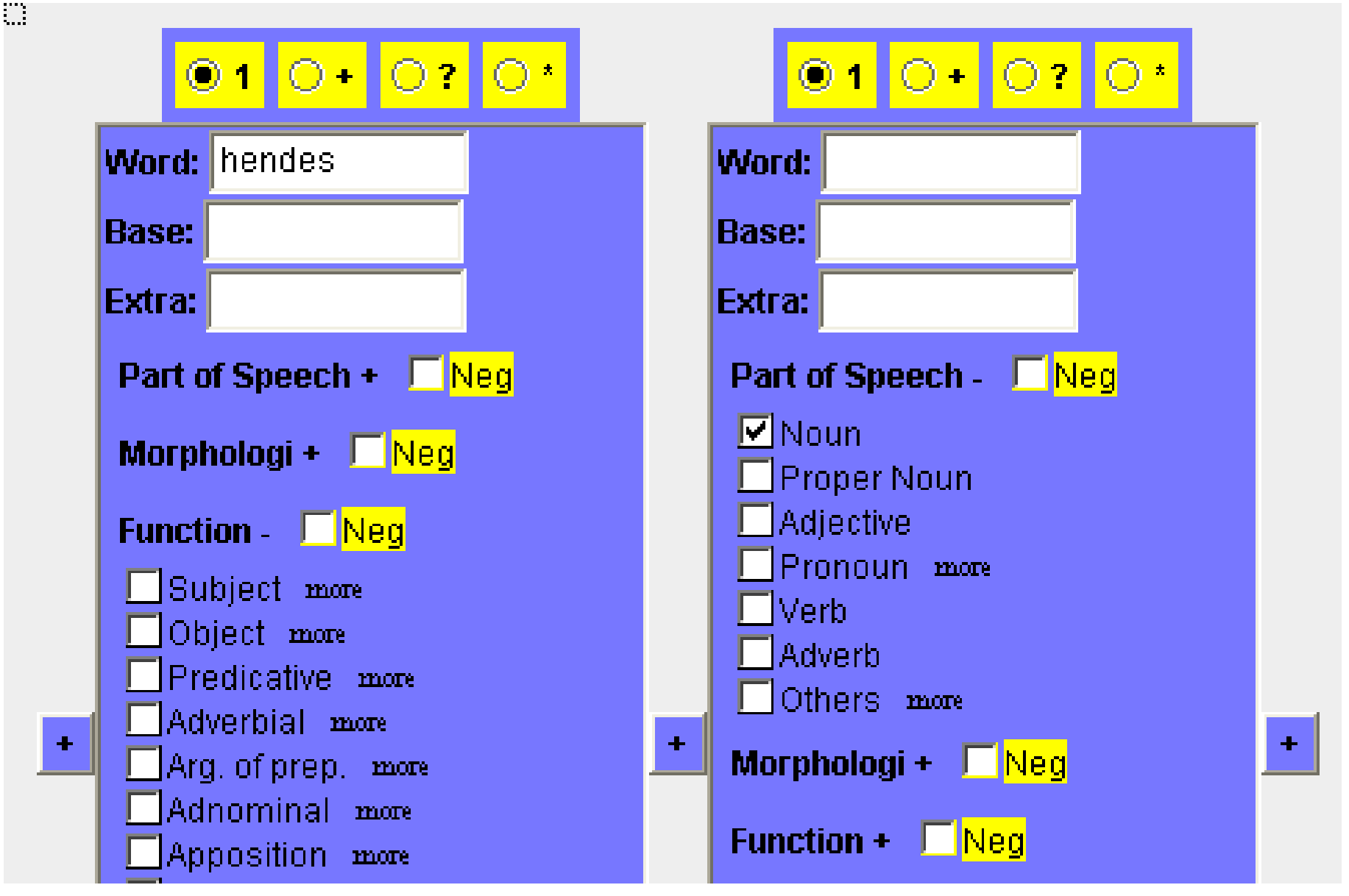
VISL credits info copyright publications links

Simple text searches: fx. eg. composita



kvadratcentimeter ibenholt og **perlemor** sidder på gribebrættet , han
de første to kvarter en række **perleafleveringer** af_sted .
Importen af **perler** og smykker steg i 1999 med 15
Som barokslottet er en **perle** af enkel pragt , er orglet i_si
En **perlerække** I_Forum havde Bob_Dylan de
ısrevyen som rigets kulturelle **perle** .
At jeg kaster **perler** for svin i et kulturelt u-land
Trods en **perlerække** af smukke melodier og fine
sbenhavn trak en to meter lang **perlekæde** ud_af sin endetarm .
, og alt af værdi , juveler , **perler** - alt .
I_går stod hun for en **perlesmykket** Maria_Stuart og en itali
En lille **perle** af en scene var , da Kelly førs
broderede fåreskinds-pelse og **perlekæder** .
: hjem og stiller bilen på den **perlegrusdækkede** gårdsplads .
etning blot er den sidste i en **perlerække** af gamle familieforretning
gudeladt lyder det dernede fra **perlegruset** .
land , hvorefter vi fangede en **perlehøne** i luften , og til_sidst fan
mulighed for at gense en række **perler** .
som ærkeenglen Gabriel var en **perlende** og rendyrket fornøjelse , en
: , Kaj-bøger , hendes elskede **perlekæder** , og hvem tog babyalarmen
på højde med den motormæssige **perle** : boksramme i stål med alu-bags

Menu based category search



Output: "raw" concordance

sort freq rel

By: Left Context / Right Context / Left Edge / Right Edge

Offset 0

Freq items 100

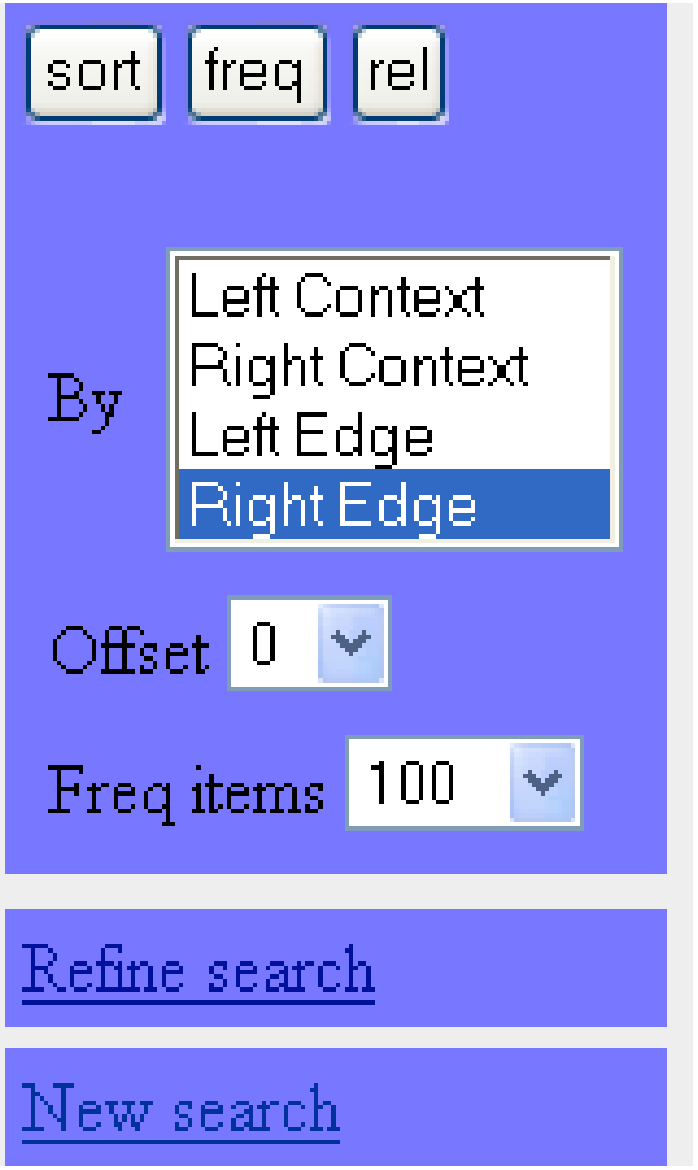
Refine search

New search

Searched for: [word="hendes"] [pos="((.*)?(N(.*)?)|((.*)?PROP)(.*)?)"]
In corpus: DAN_C90 DAN_EUROPARL
Found 12829 results (10880 1949).
1 - 50 next

INFO		¤ Da **hendes**	**mormor** døde i nc
INFO		¤ Og i betragtning af , hvad vi , **hendes**	**publikum** , har f
INFO	afstemmer de ansøgerens mål og planer med hans ¤ **hendes**		**kvalifikationer**
INFO	¤ Hun arbejdede i receptionen på et hotel , hvor **hendes**		**datter** Belinda c
INFO		¤ Og det viser sig , at **hendes**	**læge** har sagt ,
INFO		¤ Deraf **hendes**	**personligheds** al
INFO	handler om en pige , der hedder Charlie , og om **hendes**		**onkel** , der unde
INFO	han ikke er bange for heste , er stutteriet **hendes**		**afdeling** , for h
INFO	er forelsket i hende , Hortensio , der vil have **hendes**		**penge** og Gremio
INFO	¤ Hverken med tvang eller frivilligt , siger **hendes**		**bedstefar** , Poul
INFO	¤ Formelt , fordi parlamentet havde forkastet **hendes**		**plan** om prisstig
INFO		¤ Nu er **hendes**	**dominans** brudt ,
INFO	¤ I_forvejen har sagen betydet , at **hendes**		**far** for_eksempel
INFO		¤ Og da rotterne legede i **hendes**	**have** , reagerede
INFO	død 24. marts sidste år , blev bekræftet af **hendes**		**forældre** :
INFO	om pigens behov for medicin og om rigtigheden i **hendes**		**forklaring** om ep
INFO	tog han sig livrem af og viklede den to gange om **hendes**		**hals** og trak til
INFO	, løb kollegaen hen til hende og greb fast i **hendes**		**ben** .
INFO	godt , at der ikke stilles for store krav til **hendes**		**sexualitet** .

Sortering and statistics



Search
In corp
Found
1 - 50

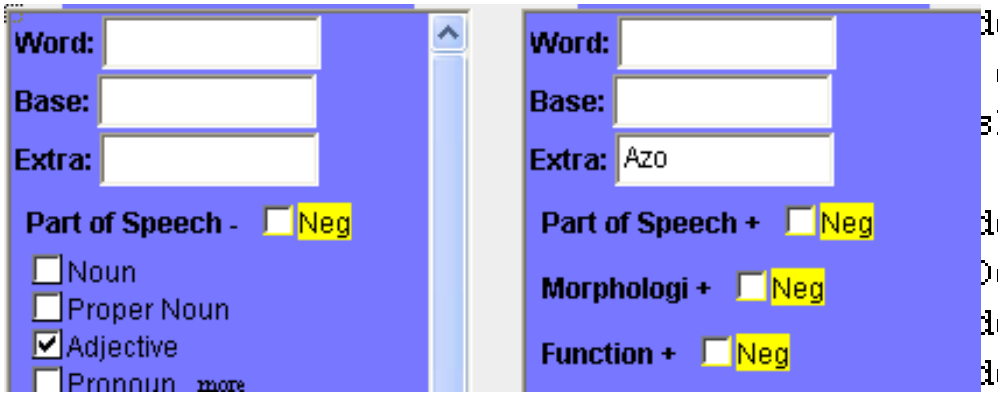
INFO
INFO
INFO afstemme
INFO ¤ Hun ar
INFO
INFO
INFO handler
INFO han
INFO er fore.
INFO ¤ Hv

imperatives

DAN_C90 (53621)

frequencies:

jf.
jfr.
Lad
lad
Læg
Tag
Hæld
Rør
Skær
Se
Sæt
jvf.
Kog
Tænk
Prøv
Jf.
Smag
Husk



animal expressions

Den nu **fire-årige hanbjørn** er ke
aget med en **gigantisk hjort** tøvende i
rer til den **politiske ræv** , fordi det
urypris med **tilhørende sølvbjørn** til
længere en **skræmt hjort** fanget i for
vivl om den **amerikanske tigers** holdba
kan få den **russiske bjørn** til at gun
ru og store **stygge ulv** .
nmarks mest **berømte løve** på en sokkel
akt med den **indre abe** er i_hvert_fald
ler er nogle **store frøer** nede i skoven
ølge med de **unge løver** , der vil køre
t_par af de **dødfødte aber** lyste fluor
erti er en **grøn marekat** en sand gour
dt le , den **vojvodinske dræven** " , üd
den **olympiske vildhest** 49er .
de **unge skakløvers** forslag b
ske **unge løver** i Venstre .
Lodne plysbjørne , der du
den **hvide hun-ulv** hjemme i Køl
Den **unge venstreløve** taler me
den **afskyelige tiger** viser si
den **russiske bjørn** som vinder

CORPUS EXERCISE TASKS (e.g. http://corp.hum.sdu.dk):

General familiarity with corpus searches:

- Regular expressions:
  - Find the longest Portuguese word!
  - Find words with anti- (with or without hyphen?)
- Lexicography:
  - Aids, aids, Sida, sida - what is "normal", and where? Frequency?
  - Empirical gender of: *personagem (a/uma @>N personagem vs. o/um @>N personagem)*
- Syntax:
  - Find subjects to the right of their verbs! Are certain verbs more likely to occur in these constructions than others?
  - Find a noun phrase with as many dependents as possible (Cqp or treebank)
- Find a relative clause within a relative clause! (DN+fcl << (DN+fcl << (DN+fcl << DN+fcl))
- Semantics:
  - Find male/femal-typical nouns: N de @N< ele/ela @P< (Brazilian data: folha!)
  - Find time adverbs and other time expressions and classify them! (high level task)
  - Find profession nouns, using different methods (suffix, context, special tags: Hprof in Público)