Syntactic analysis of natural languages

Vojtěch Kovář

NLP Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno

xkovar3@fi.muni.cz

PA153 Natural Language Processing

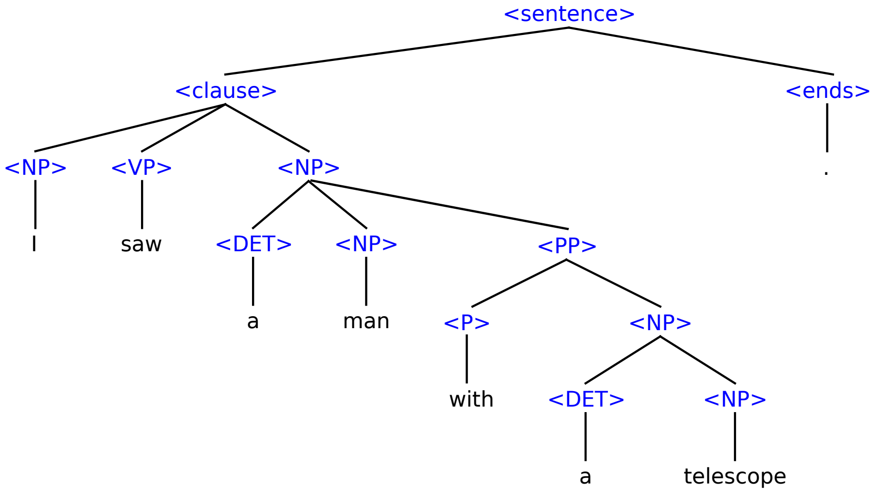
Syntactic analysis

- What?
  - reveal the structure of the sentence
  - relationships among words, phrases
- Why?
  - basis for more informed language analysis
  - more than keywords
  - semantic and logical analysis, question answering, ...
  - applications can benefit from syntactic information
  - red brick house vs. red house brick vs. brick house red
- Origins
  - Noam Chomsky: Syntactic structures (1957)
  - theory of formal languages

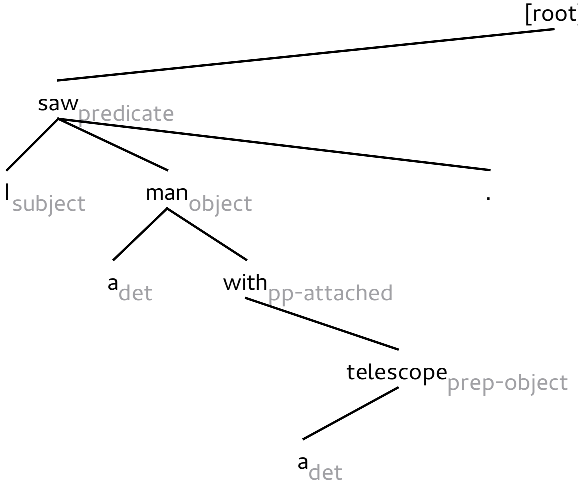
Automatic syntactic analysis of natural languages

- Preprocessing
  - sentence boundary detection
  - word segmentation
  - morphological analysis and disambiguation
  - (named entity MWE recognition, lexical semantics, ...)
  - compatibility issues
- Encoding
  - phrase structure formalism
  - dependency formalism
  - partial analysis
  - advanced – CCG, LFG, HPSG, TAG

Phrase structure formalism – example



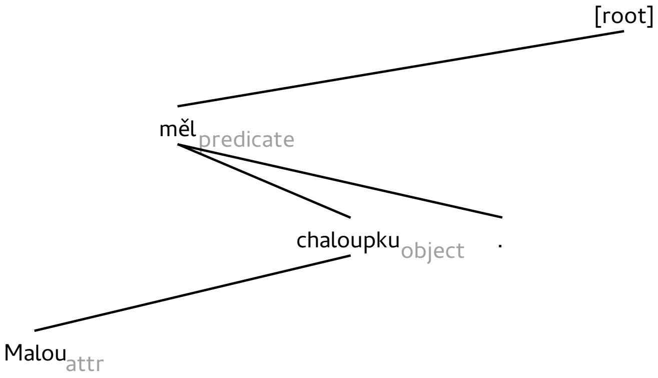
Dependency formalism – example



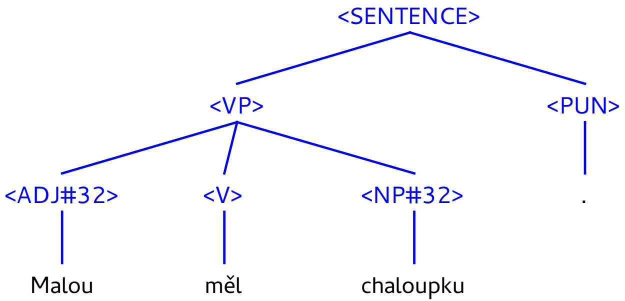
Dependency vs. phrase-structure

- Non-projectivity
  - disconnected phrases
  - not natural in the phrase structure notation
  - 20% of Czech sentences are reported to contain a non-projective dependency
- Phrase structure – more fine-grained analysis
  - (new (queen of beauty))
  - (new generation)(of fighters)
- Coordinations and other "flat" phenomena
  - not natural in the dependency notation
  - problem for dependency analysis

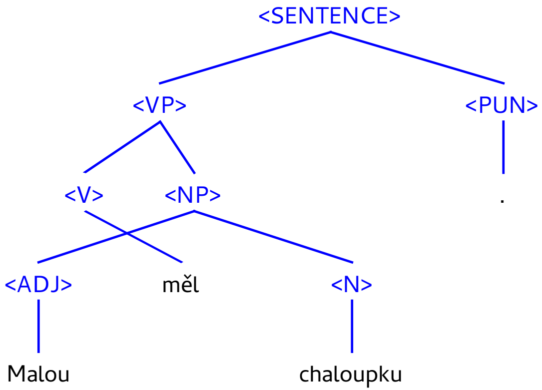
Non-projectivity – example



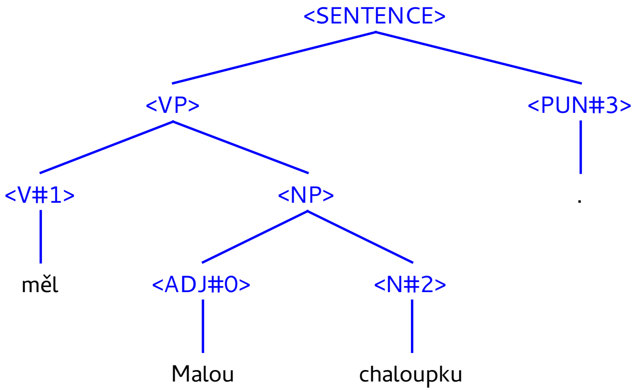
Non-projectivity in phrase structure formalism



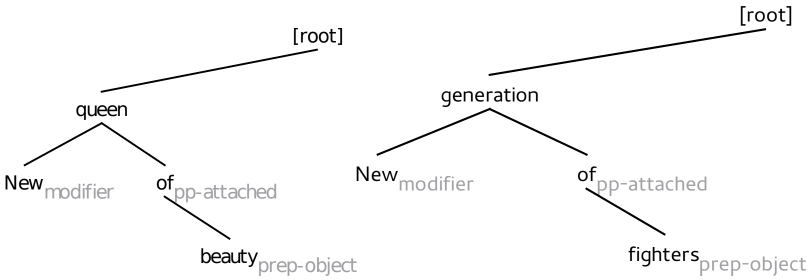
Non-projectivity in phrase structure formalism



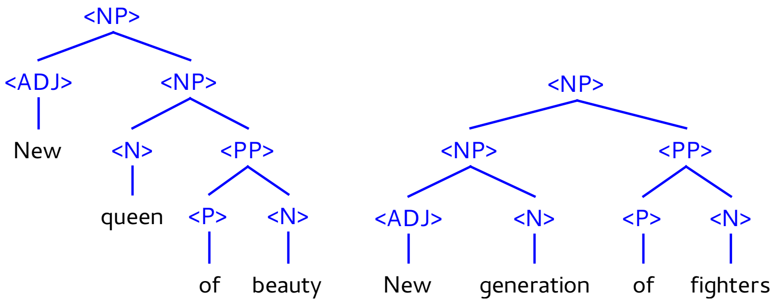
Non-projectivity in phrase structure formalism



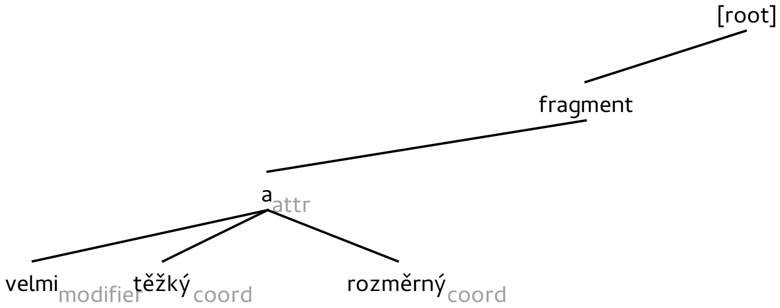
Phrase structure expressivity



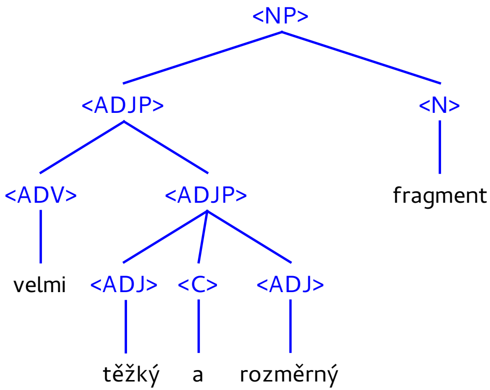
Phrase structure expressivity



Coordinations – dependency structure



Coordinations – phrase structure

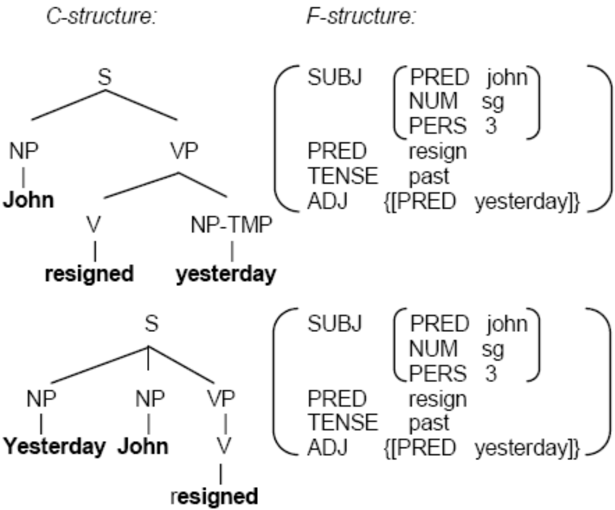


CCG: Combinatorial Categorial Grammar

$$the : NP/N \qquad dog : N \qquad John : NP \qquad bit : (S\backslash NP)/NP$$

$$
\dfrac{
\dfrac{\dfrac{the}{NP/N} \quad \dfrac{dog}{N}}{NP}{>}
\qquad
\dfrac{\dfrac{bit}{(S\backslash NP)/NP} \quad \dfrac{John}{NP}}{S\backslash NP}{>}
}{S}{<}
$$

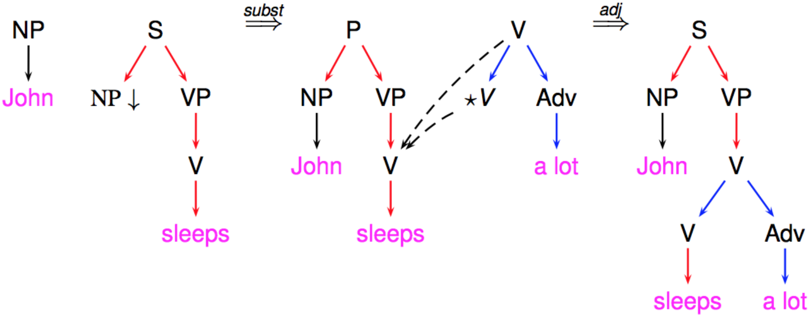
LFG: Lexical Functional Grammar



HPSG: Head-driven Phrase Structure Grammar

$$\begin{bmatrix} \textit{word} & \\ \text{PHON} & \langle \text{`walks'} \rangle \\ \text{SYNSEM} & \begin{bmatrix} \textit{synsem} & \\ \text{CAT} & \begin{bmatrix} \textit{category} & \\ \text{HEAD} & \textit{verb} \\ \text{VALENCE} & \begin{bmatrix} \text{SUBJ} & \left\langle \begin{bmatrix} \textit{synsem} & \\ \text{CAT}\,|\,\text{HEAD} & \textit{noun} \\ \text{CONT} & \boxed{1} \begin{bmatrix} \textit{ref-index} & \\ \text{PER} & \textit{3rd} \\ \text{NUM} & \textit{sing} \end{bmatrix} \end{bmatrix} \right\rangle \\ \text{COMP} & \langle \rangle \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \textit{content} & \\ \text{WALKER} & \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

TAG: Tree Adjoining Grammar



Parsing methods

- Rule-based
  - RASP, synt, SET, Žabokrtský, Dis/VaDis
  - manually created grammar
  - CFG (CKY parser, chart parser), dependency grammar, Prolog DCG, ...
- Statistical
  - MaltParser, MST Parser, Stanford parser, ...
  - grammars created from annotated data by statistical methods
  - direct guessing the tree shape

Parsing evaluation

- Treebanks
  - corpora manually annotated for syntactic structure
  - Penn Treebank, Prague Dependency Treebank (PDT)
- Tree similarity metrics
  - PARSEVAL: precision, recall, F-score over phrases
  - Leaf-ancestor assessment: edit distance over root-leaf paths
  - dependency precision
  - labelled or unlabelled
  - best results: 85–93 percent

Problems

- Central problem
  - massive ambiguity
  - "I saw a man with a telescope"
  - "A plane fell into a field next to a forest."
  - problems with evaluation
- Is the task well-defined?
  - inter-annotator agreement rarely reported
  - in case of PDT around 90%
  - Sampson showed that above 95% is unreachable
  - $\rightarrow$ current parsers are very good
  - however, rather low usage in applications

Problems (II)

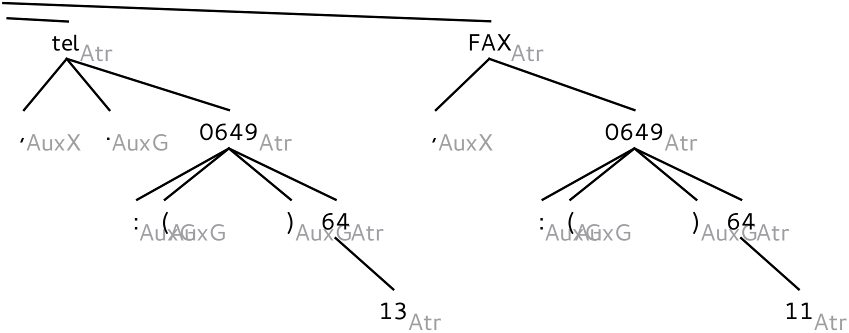
- Low usage
  - compared to e.g. morphological tagging
  - no use in Google, Seznam, Facebook, ...
  - Wikipedia page for information extraction does not even mention parsing or syntax
  - neither does a Czech question answering system (Konopík, Rohlík)
  - ACL anthology: 7,232 matches for word "parser", 133 matches for using parsers (Jakubíček)
- Are the results useless?

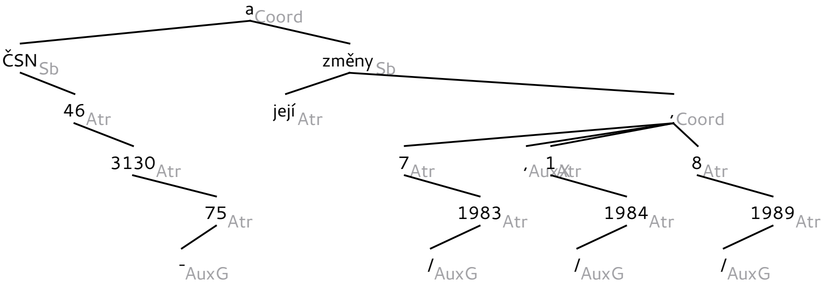
Problems (III)

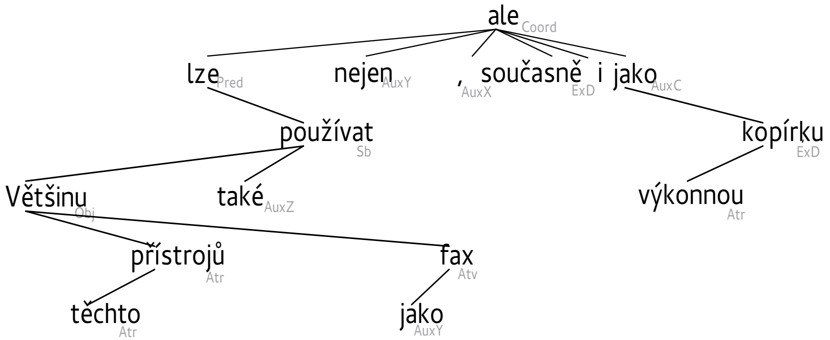
- Application-sparse output
  - trees do not provide all the information needed
  - but at the same time they do contain noise
- Application-free evaluation
  - tree similarity metrics do not correlate well with accuracy of the end applications
  - as illustrated by Myiao, Google research, our collocation extraction research
- Technical aspects
  - parsers hard to run, output not readable

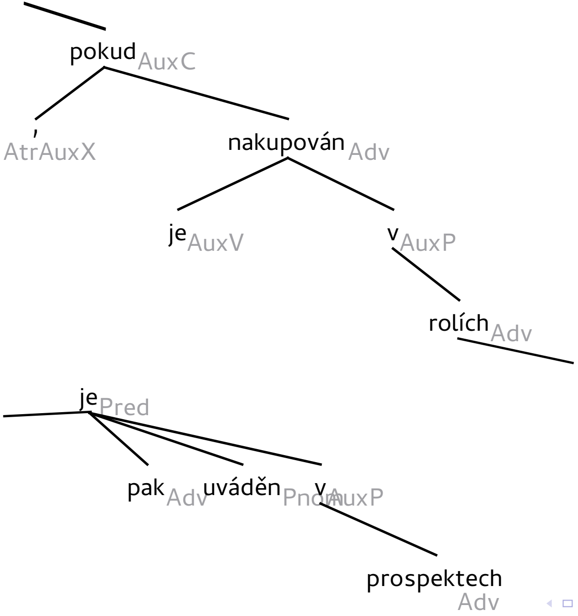
Treebank problems

- Apart from evaluation problems, treebanks are
    - expensive
    - old
    - domain-specific
    - unambiguous
- Treebank formalisms enforce
    - annotation manuals containing hundreds of pages
    - senseless annotations and garbage


tel Atr
FAX Atr
, AuxX
. AuxG
0649 Atr
, AuxX
0649 Atr
: AuxX
(AuxG
) AuxG
64 Atr
: AuxX
(AuxG
) AuxG
64 Atr
13 Atr
11 Atr


a Coord
ČSN Sb
změny Sb
46 Atr
její Atr
, Coord
3130 Atr
7 Atr
, AuxX
1 Atr
8 Atr
75 Atr
1983 Atr
1984 Atr
1989 Atr
- AuxG
/ AuxG
/ AuxG
/ AuxG


ale Coord
lze Pred
nejen AuxY
, AuxX
současně ExD
i
jako AuxC
používat Sb
kopírku ExD
Většinu Obj
také AuxZ
výkonnou Atr
přístrojů Atr
fax Atv
těchto Atr
jako AuxY


pokud AuxC
,
AtrAuxX
nakupován Adv
je AuxV
v AuxP
rolích Adv
je Pred
pak Adv
uváděn Pnom
v AuxP
prospektech
Adv

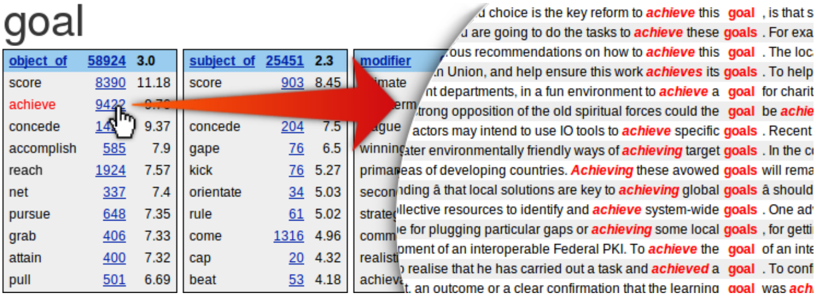
Proposed solution: You aren't gonna need it

- Rapid application development
  - „worse is better"
  - „keep it simple stupid" (KISS)
  - „you aren't gonna need it" (YAGNI)
  - completeness, consistency, correctness, simplicity
- Implications
  - start from applications
  - strong emphasis on interaction with applications
  - do not develop/implement theory that is not immediately needed
  - simple, imperfect parsers, possibly task-specific
  - rule based first, until we find what we actually need
  - extrinsic evaluations

Sketch grammar: A shallow approach to syntax

- Designed for collocation extraction
  - Kilgarriff and Rychlý, The Sketch Engine
  - based on Corpus Query Language
  - results of queries scored statistically
  - → pragmatic partial syntactic analysis
- Extensions
  - multi-word sketches
  - bilingual word sketches
  - terminology extraction
  - bilingual terminology extraction

Word Sketch – original



Sketch grammar example

```
*DUAL
=subject/subject_of

    2:[tag="N.*"]  [tag="RB.?"]{0,3}  [lemma="be"]?
       [tag="RB.?"]{0,2}  1:["V.[^N]?"]
```

Terminology extraction

Term	Frequency	Freq/mill	Score
carbon dioxide	373	3864.3	37.5
global warming	317	3284.1	30.8
water vapor	71	735.6	8.3
greenhouse effect	69	714.8	8.1
greenhouse gas	71	735.6	8.0
climate change	78	808.1	7.6
industrial ecology	27	279.7	3.8
fossil fuel	26	269.4	3.6
surface temperature	20	207.2	3.1
carbon cycle	19	196.8	3.0

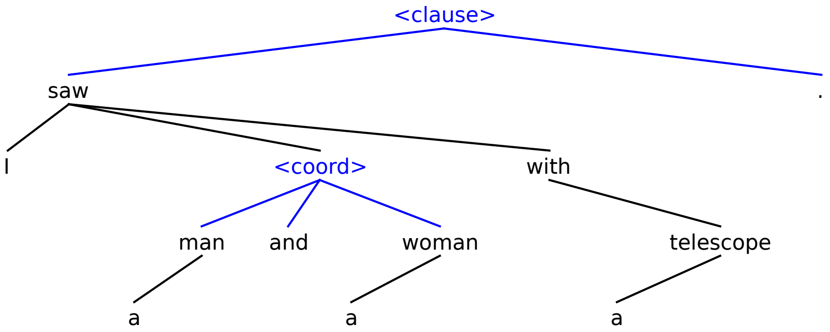
Sketch grammar for terminology extraction

```
=terms
*COLLOC "%(2.lc)_%(1.lc)"

  2:[tag=="NN" | tag=="JJ" | tag=="VVG"]  1:[tag=="NN"]

*COLLOC "%(3.lc)_%(2.lc)_%(1.lc)"

  3:[tag=="NN" | tag=="JJ" | tag=="VVG"]
        2:[tag=="NN" | tag=="JJ" | tag=="VVG"]
        1:[tag=="NN"]
```

SET – a light-weight parsing system

- Hybrid trees
  - combination of dependency and phrase structure formalisms
  - readability, natural analysis
- Pattern matching grammar
  - similar to CQL
  - manually created and ranked rules
  - rules → matches → sorting → best tree

Hybrid tree



SET rule example

```
TMPL: (tag k5)  ...  $AND  ...  (tag k5)
      MARK 0 2 4 <coord>  PROB 500  HEAD 2
$AND(word): , a ani nebo
```

Synt – a traditional CFG+ parser

- CFG backbone + contextual actions
    - manually created CFG grammars for Czech, Slovak, English
    - statistical ranking of rules
    - chart parser + extensions

Conclusions

- There are many ways to approach syntactic analysis
  - none of them became dominant in practice (yet?)
- Basic formalisms
  - dependencies
  - phrase structure
- Manual as well as statistical approaches

Links

www.diotavelli.net/people/void/demos/cky.html
en.wikipedia.org/wiki/Definite_clause_grammar
en.wikipedia.org/wiki/Combinatory_categorial_grammar
en.wikipedia.org/wiki/Head-driven_phrase_structure_grammar
nlp.fi.muni.cz/projekty/wwwsynt
nlp.fi.muni.cz/projekty/wwwsynt/query.cgi
nlp.fi.muni.cz/trac/set
nlp.fi.muni.cz/projekty/set/wwwset.cgi/first_page
ufal.mff.cuni.cz/pdt2.0/index-cz.html