Automatic syntactic analysis for real-world applications

Vojtěch Kovář

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

xkovar3@fi.muni.cz

Outline

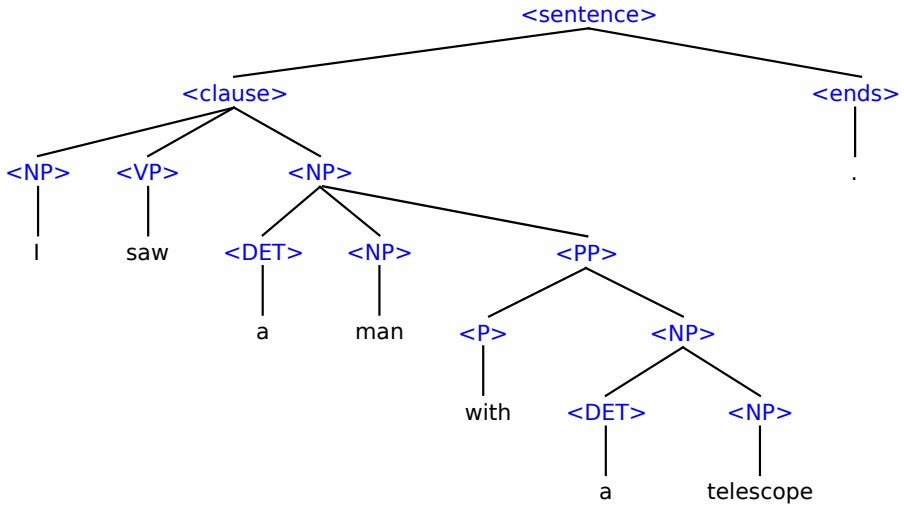
Challenges in natural language processing

- Information retrieval
- Information extraction
- Question answering
- Automatic reasoning – textual entailment
- Authorship recognition
- Grammar checking
- Collocation extraction
- Terminology extraction
- Hidden applications
  - morphology disambiguation
  - anaphora resolution
  - automatic extraction of semantic frames
  - extraction of lexical semantic information
  - natural language generation

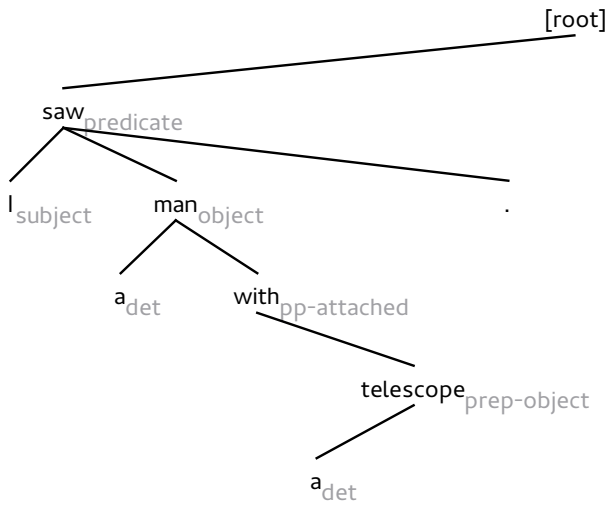
Automatic syntactic analysis of natural languages

- Revealing the sentence structure
- Preprocessing
  - sentence boundary detection
  - word segmentation
  - morphological analysis and disambiguation
  - named entity & MWE recognition, lexical semantics, ...
- Encoding
  - phrase structure formalism
  - dependency formalism
  - partial analysis
  - advanced – CCG, HPSG, TAG, LFS

Phrase structure formalism – example



Dependency formalism – example



Dependency vs. phrase-structure

- Non-projectivity
  - disconnected phrases
  - not natural in the phrase structure notation
- 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

Parsing methods

- Rule-based
  - set of rules (CFG, pattern-matching, ...)
  - RASP, synt, SET, Žabokrtský, Dis/VaDis
- Statistical
  - models learned from annotated data
  - MaltParser, MST Parser, Stanford parser, ...

State of the art 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: 85–90 percent

Criticism of state of the art (I)

- Is the task well-defined?
  - inter-annotator agreement rarely reported
  - in case of PDT around 90%
  - Sampson showed that above 95% is unreachable
  - → current parsers are very good
- Low usage
  - compared to e.g. morphological tagging
  - are the results useless?

Marie Mikulová and Jan Štěpánek. Annotation procedure in building the Prague Czech-English dependency treebank.

Geoffrey Sampson and Anna Babarczy. Definitional and human constraints on structural annotation of English.

Criticism of state of the art (II)

- 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
- Technical aspects
  - parsers hard to run, output not readable

Yusuke Miyao, Kenji Sagae, Rune Sætre, Takuya Matsuzaki, and Jun'ichi Tsujii. Evaluating contributions of natural language parsers to protein–protein interaction extraction.
Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, Masakazu Seno, and Hideto Kazawa. Training a parser for machine translation reordering.

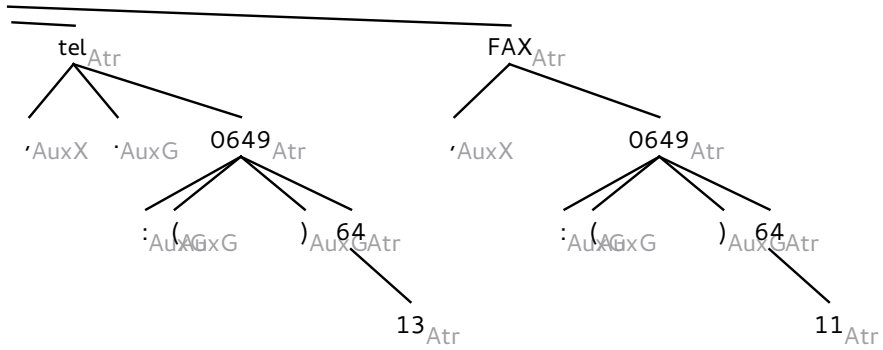
Proposed solution: You aren't gonna need it

- 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

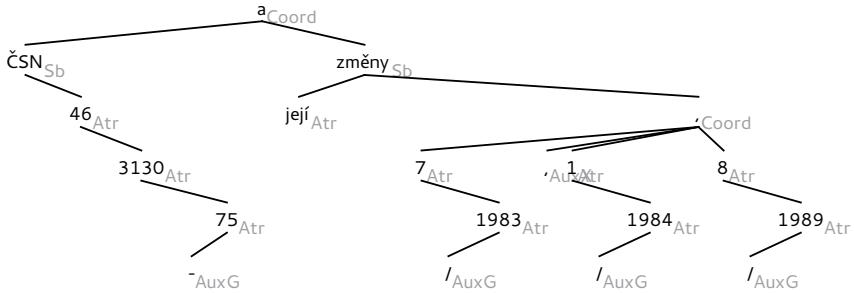
Bushbank: Alternative syntactic annotation

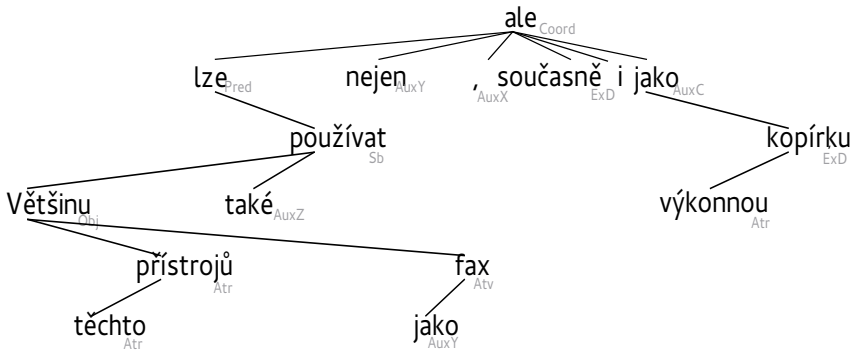
- Apart from evaluation problems, treebanks are
  - expensive
  - old
  - domain-specific
  - unambiguous
- Treebank formalisms enforce
  - annotation manuals containing hundreds of pages
  - senseless annotations and garbage

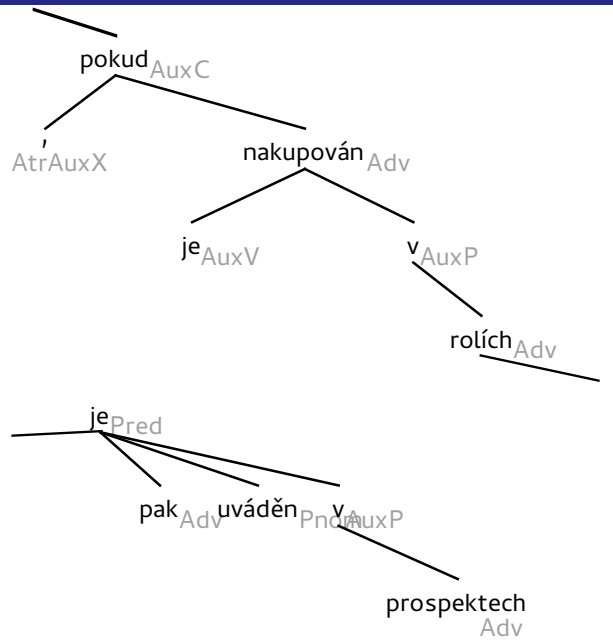
, tel.: (0649) 64 13, FAX: (0649) 64 11



ČSN 46 3130-75 a její změny 7/1983, 1/1984, 8/1989



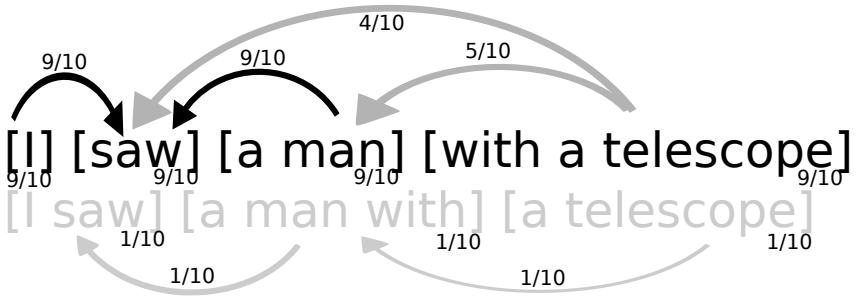




Bushbank: Alternative syntactic annotation

- No useless information
  - noun, prepositional and verb phrases
  - dependencies
  - words can be outside phrases
  - comprehensible information directly usable in applications
- Cheap
  - yes/no annotation of parser output
  - 10 times faster than treebank annotation
  - annotation manual of 5 pages (with 92% agreement)
- Natural ambiguity
  - using inter-annotator agreement

Bushbank: Alternative syntactic annotation



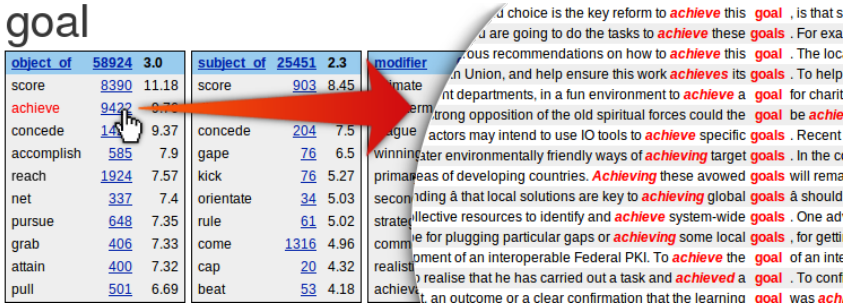
Parser evaluation against PDT and Czech Bushbank

Parser	PDT precision (%)	PHRASE F-score (%)
SET	56.0	81.4
Collins	80.9	73.0
MaltParser	85.8	49.6
MST Parser	84.7	49.7
IOBBER	N/A	**90.3**

Sketch grammar: A shallow approach to syntax

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

Word Sketch – original

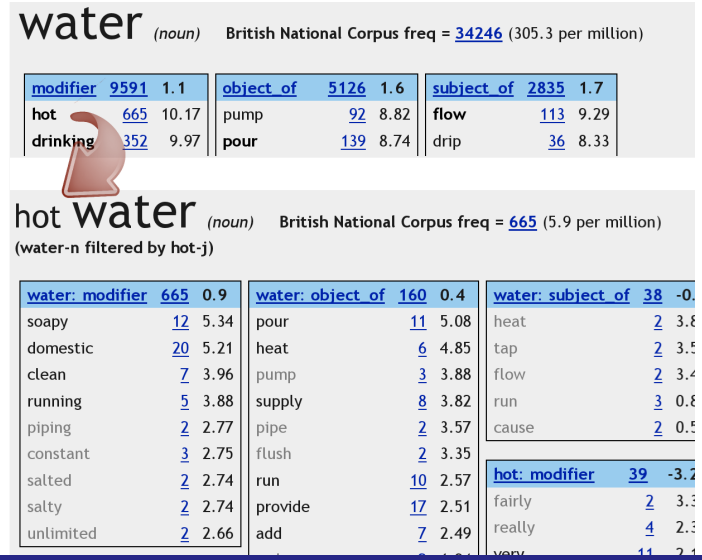


Sketch grammar example

```
*DUAL
=subject/subject_of

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


Multiword sketch



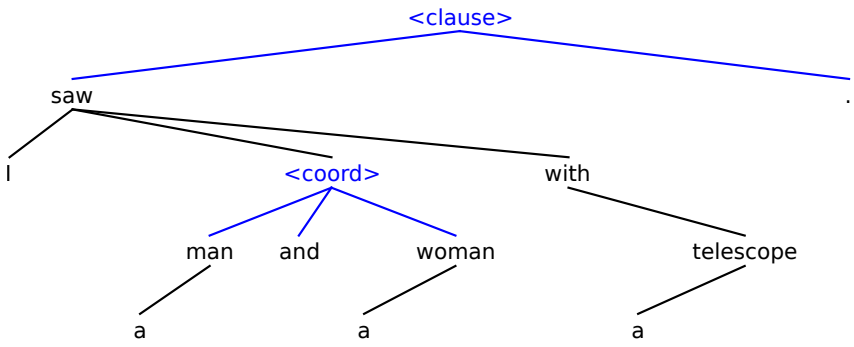
Terminology extraction

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

SET – a light-weight parsing system

- Hybrid trees
  - combination of dependency and phrase structure formalisms
  - readability, natural analysis
- Pattern matching grammar
  - similar to the Corpus Query Language
  - ranked pattern matching rules
  - rules → matches → sorting → best tree

Hybrid tree



SET rule example

```
TMPL: (tag k5)  ...  $AND  ...  (tag k5)
      MARK 0 2 4 <coord>

$AND(word): , a ani nebo
```

Applications

- Information extraction for Czech
  - SET phrases → semantic classification → facts
  - 70% accuracy
- Textual entailment for Czech
  - inference rules over SET syntactic phrases
  - 86% precision
- Authorship verification for Czech
  - Authorship Recognition Tool: machine learning
  - SET syntactic features → improvement 3–7%

Applications (II)

- Punctuation detection for Czech
  - special SET grammar
  - precision 97.1%, recall 56.8%
- Subject-predicate disagreement detection
  - modified subject rules
  - precision 100%, recall 18%
  - correct tagging → precision 100%, recall 64%
  - (small testing set)
- Collocation extraction
  - detailed evaluation of the application
  - creating gold standard data
  - word sketches for Czech from different parsers

Parser evaluation with PDT and using collocation extraction

Parser	PDT score (%)	collocation extraction F-5 (%)
Sketch grammar	N/A	60.3
Synt	N/A	54.0
SET	56.0	57.2
MST Parser	84.7	57.8
MaltParser	85.8	57.6

Applications (III)

- Terminology extraction
  - for 10 languages, evaluated on 5 languages
  - precision 67–95%
- Bilingual terminology extraction
  - preliminary evaluation on English vs. 4 other languages
  - precision 35–88%
- Automatic extraction of lexical semantics
  - Marek Grác
  - some collocations relate to specific semantic class
  - best result: SET + Sholva ontology
  - precision up to 80%, recall up to 60%, best F = 53%

Applications (IV)

- Czech phrase declension
  - Zuzana Nevěřilová
  - using SET for phrase head detection
  - accuracy 90.6%
- Anaphora resolution
  - Saara + Aara
  - precision around 40%
  - both using SET for markable detection
- Valency frame induction
  - Jiří Materna
  - corpus-driven semantic verb frames
  - frame data from SET

Applications (V)

- Ongoing applications
  - theme-rheme identification for Czech
  - intrinsic corpus evaluation with SET
  - question answering for Czech
  - syntactic information retrieval for Czech

Conclusions

- Applications prove that methodology is correct
  - our parsers are used more than state-of-the art tools
  - syntactic information brings clear advantages
  - SET is the most used Czech parser
  - application based accuracy is comparable to the state-of-the art tools
  - application based evaluations do not correlate well with treebank evaluations
- Syntactic analysis needs to be based on applications