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

1 Introduction
2 State of the art
3 Bushbank
4 Sketch grammar
5 SET parser
6 Applications
7 Conclusions

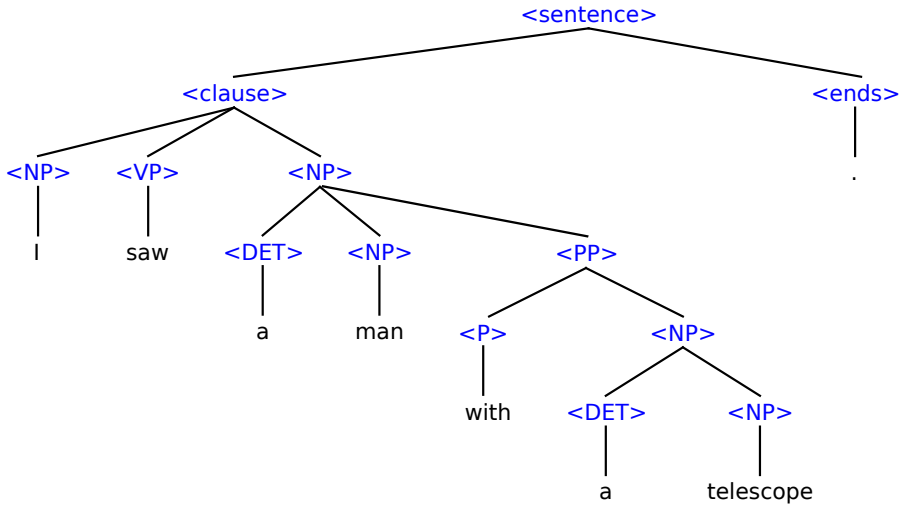
Challenges in natural language processing

- 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

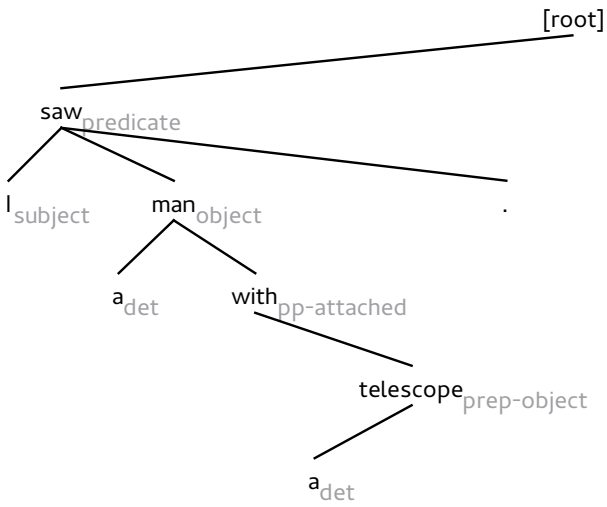
Automatic syntactic analysis of natural languages

- 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, HPSG, TAG, LFS

Phrase structure formalism – example



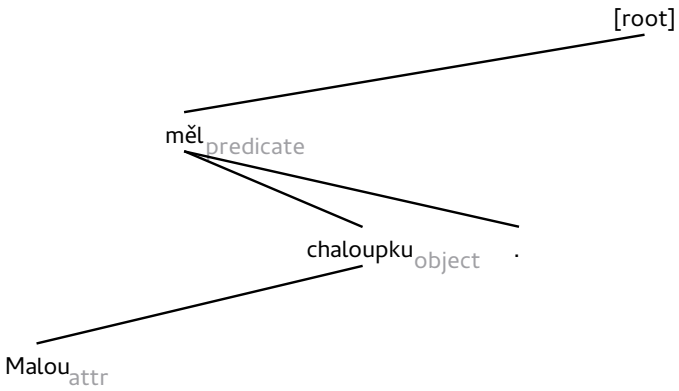
Dependency formalism – example



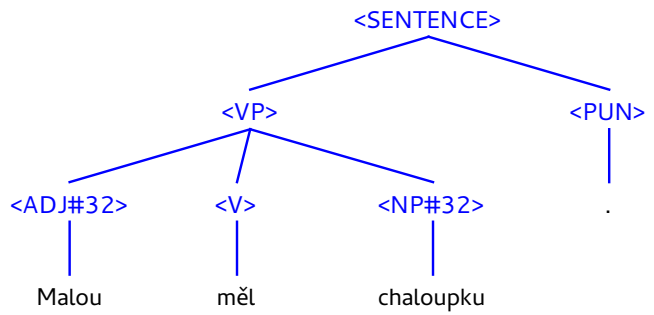
Dependency vs. phrase-structure

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

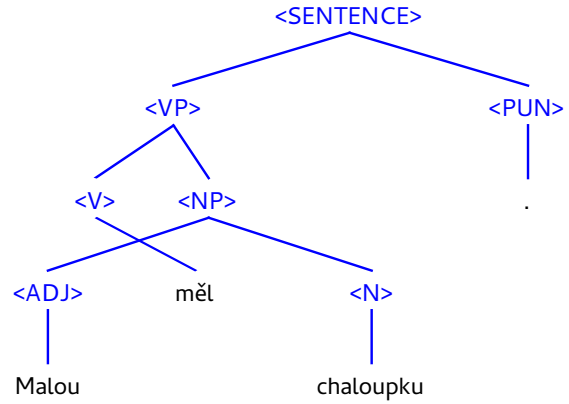
Non-projectivity – example



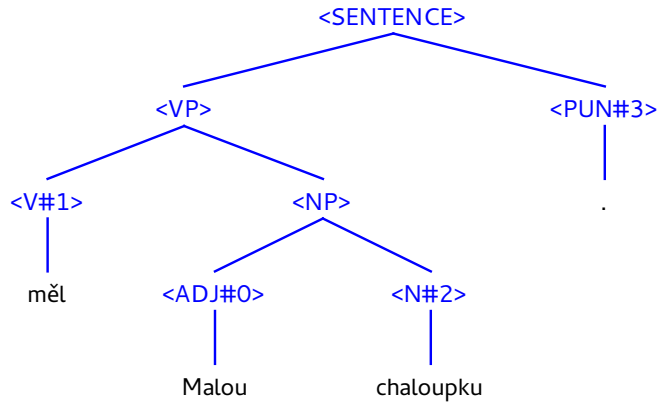
Non-projectivity in phrase structure formalism



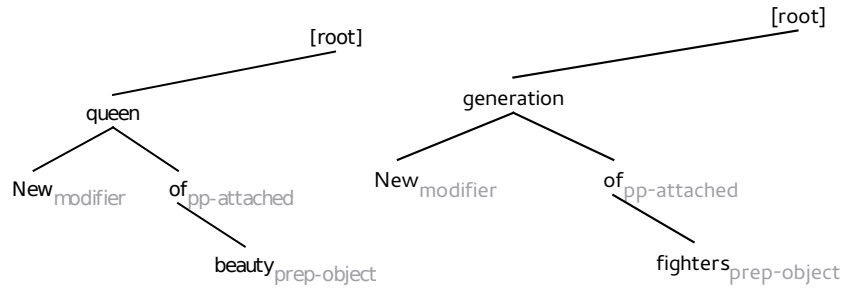
Non-projectivity in phrase structure formalism



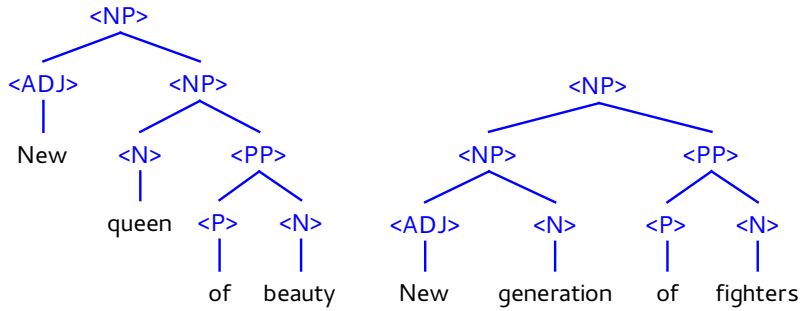
Non-projectivity in phrase structure formalism



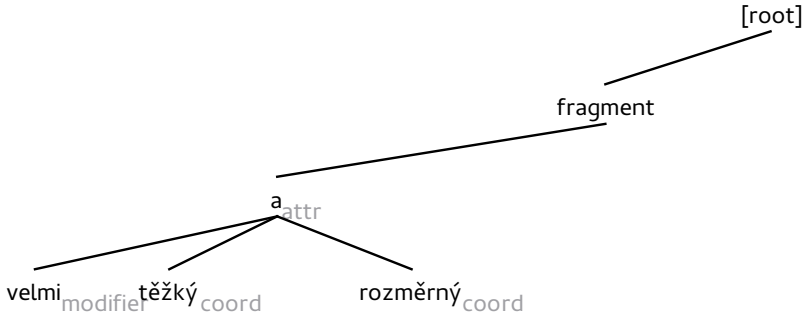
Phrase structure expressivity



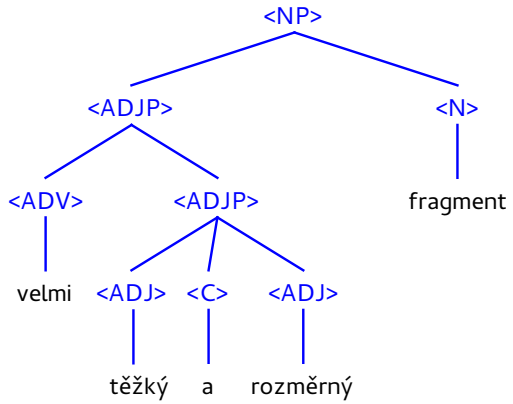
Phrase structure expressivity



Coordinations – dependency structure



Coordinations – phrase structure



Parsing methods

- Rule-based
  - RASP, synt, SET, Žabokrtský, Dis/VaDis
- Statistical
  - 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

Criticism of state of the art (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?

Criticism of state of the art (III)

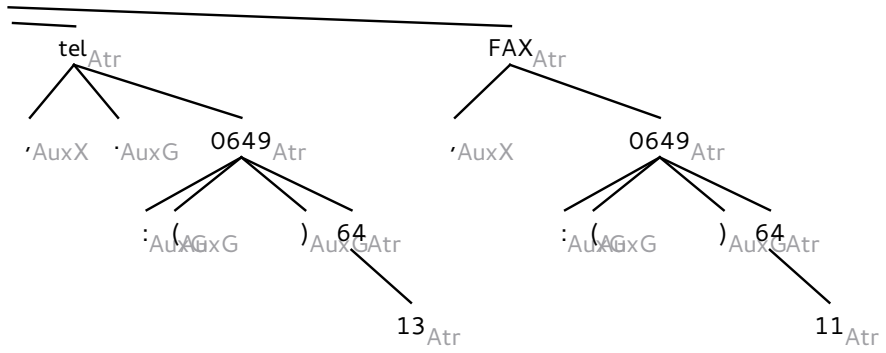
- Application-sparse output
  - trees do not provide all the information needed
  - but at the same time they do contain noise
- Application-free evaluation
  - tree similarity metrics do not correlate well with accuracy of the end applications
  - as illustrated by Myiao, Google research, our collocation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

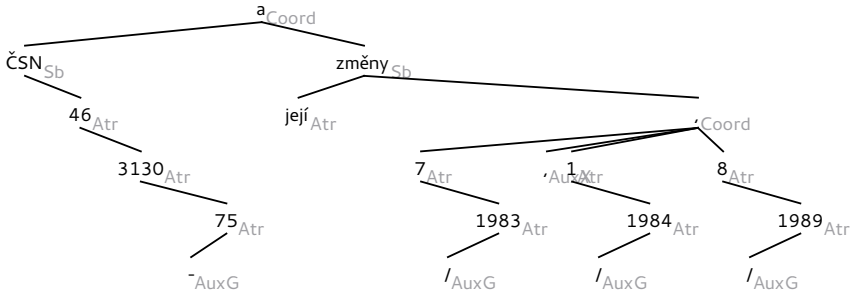
Proposed solution: You aren't gonna need it

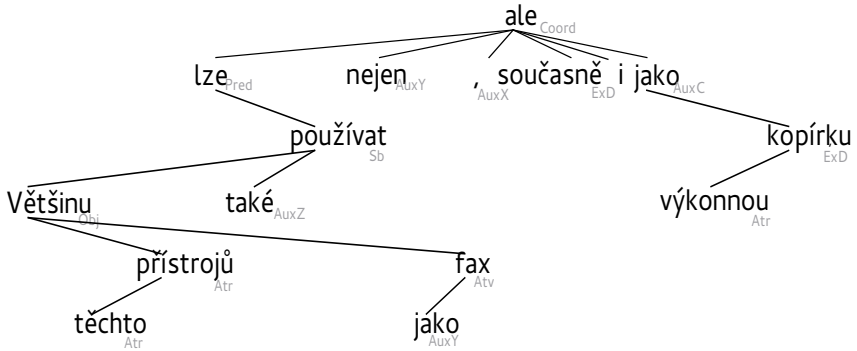
- Rapid application development
  - „worse is better"
  - „keep it simple stupid" (KISS)
  - „you aren't gonna need it" (YAGNI)
  - completeness, consistency, correctness, simplicity
- Implications
  - start from applications
  - strong emphasis on interaction with applications
  - do not develop/implement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

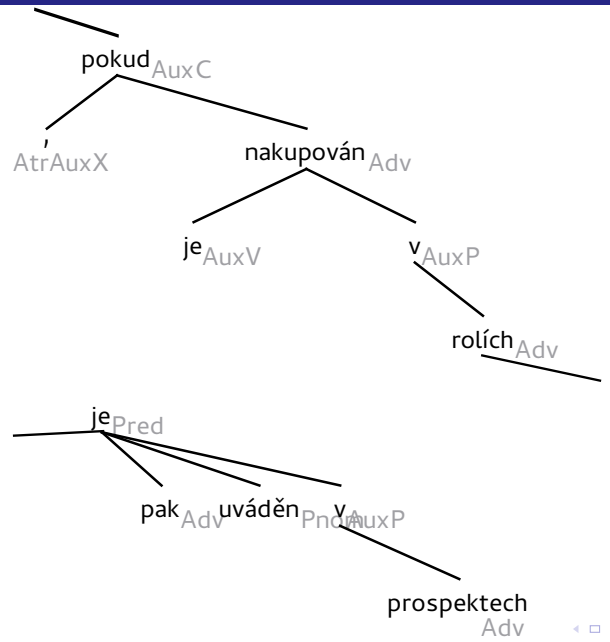
Bushbank: Alternative syntactic annotation

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





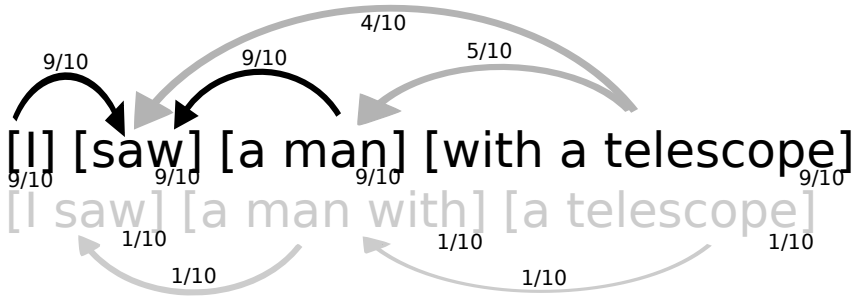




Bushbank: Alternative syntactic annotation

- 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



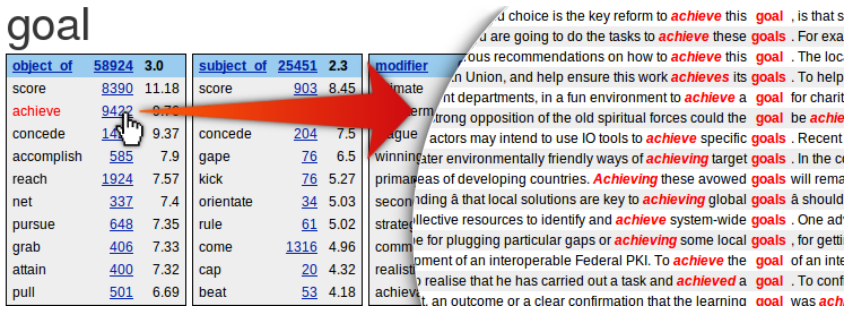
Parser evaluation against PDT and Czech Bushbank

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
  - 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]?"]
```

Multiword sketch

water *(noun)* British National Corpus freq = 34246 (305.3 per million)

modifier	9591	1.1	object_of	5126	1.6	subject_of	2835	1.7
hot	665	10.17	pump	92	8.82	flow	113	9.29
drinking	352	9.97	pour	139	8.74	drip	36	8.33

hot water *(noun)* British National Corpus freq = 665 (5.9 per million)
(water-n filtered by hot-j)

water: modifier	665	0.9	water: object_of	160	0.4	water: subject_of	38	-0.
soapy	12	5.34	pour	11	5.08	heat	2	3.8
domestic	20	5.21	heat	6	4.85	tap	2	3.5
clean	7	3.96	pump	3	3.88	flow	2	3.4
running	5	3.88	supply	8	3.82	run	3	0.8
piping	2	2.77	pipe	2	3.57	cause	2	0.5
constant	3	2.75	flush	2	3.35			
salted	2	2.74	run	10	2.57	hot: modifier	39	-3.2
salty	2	2.74	provide	17	2.51	fairly	2	3.3
unlimited	2	2.66	add	7	2.49	really	4	2.3

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

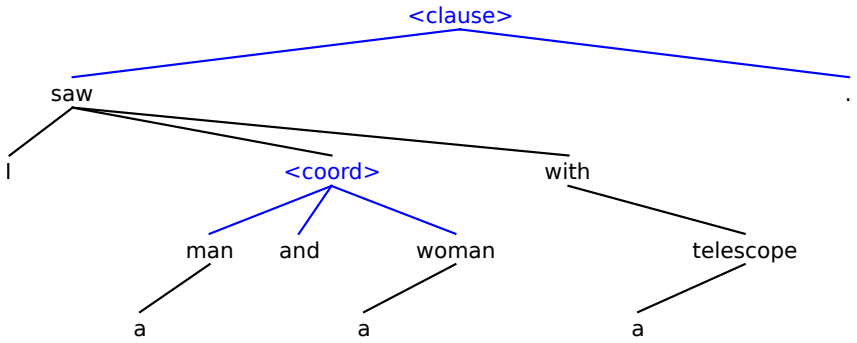
Sketch grammar for terminology extraction

```
=terms
*COLLOC "%(2.1c)_%(1.1c)"

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

*COLLOC "%(3.1c)_%(2.1c)_%(1.1c)"

  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
  - 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
```


Applications

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

Automatic valency frames

jíst$_1^{impf}$
-frame: **AG**<person:1>$^{obl}_{kdo1}$ **VERB**^{obl} **SUBS**<food:1>$^{obl}_{co4}$ **INS**<cutlery:2>$^{opt}_{čím7}$

JÍST

0.263666
frame 902

SUBJECT		ACC_OBJECT	
46		256	
0.078716	člověk	0.052979	pivo
0.06375	rodič	0.033961	alkohol
0.035257	žena	0.031826	jídlo
0.033386	matka	0.026975	voda
0.029069	otec	0.024646	krev
0.022305	muž	0.021929	cigareta

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