Extracting Phrases from PDT 2.0

Vašek Němčík
(xnemcik@fi.muni.cz)

NLP Centre
FI MU Brno

RASLAN
December 2, 2011

Outline

1. Introduction and motivation
2. Examples
3. Output format
4. Technical details

Motivation

- PDT contains *interesting* data
- The data exhibits a complex structure
- It is not easy to use the data

$\Longrightarrow$ PDT in a ready-to-use format

PDT Data Structure

- Theoretical background (FGD)
- Multiple layers
- Dependency trees
- A rich set of attributes

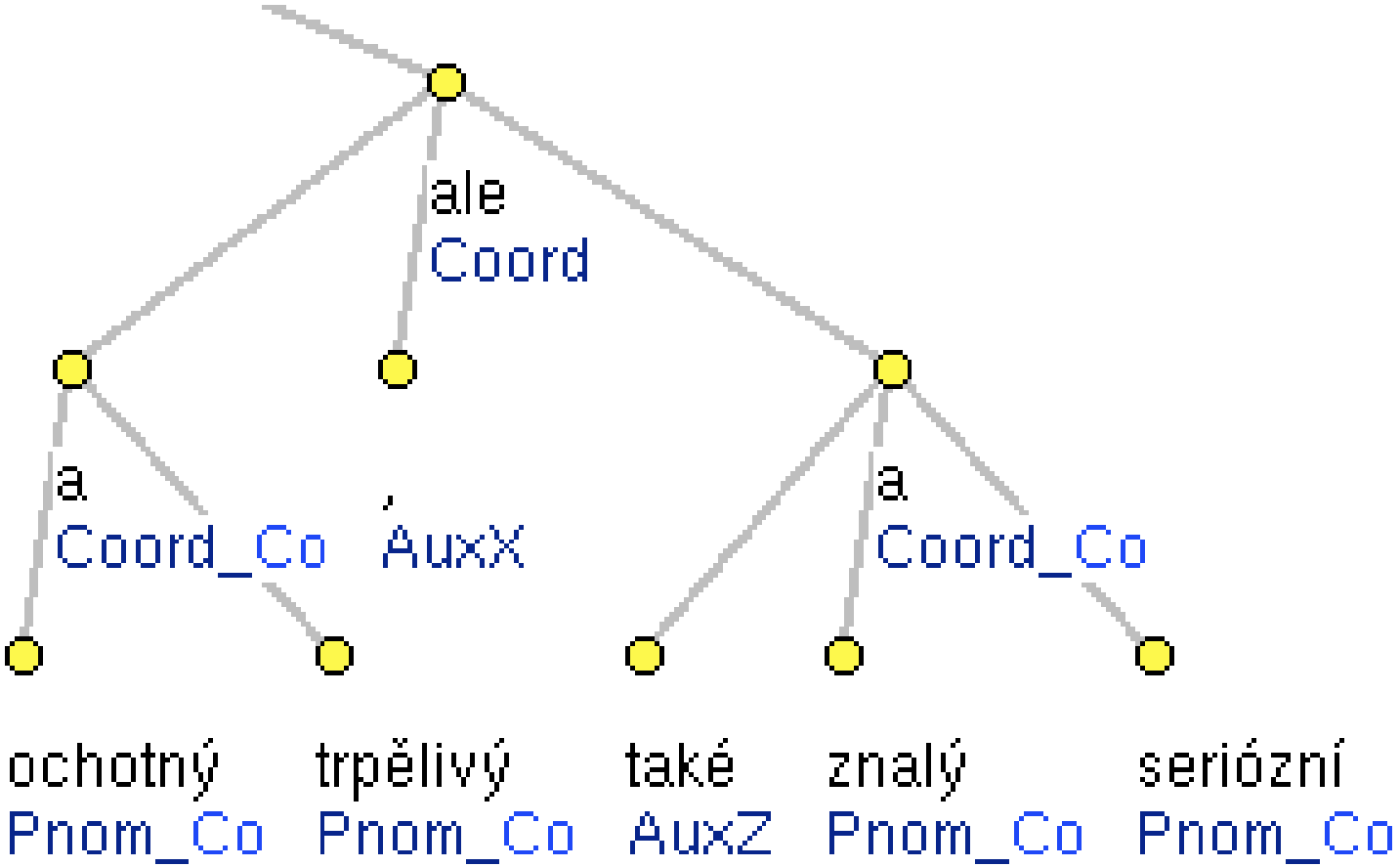
Remarkable Constructions

- Coordinations
  - a different edge type
- Shared Attributes
  - they break the tree hierarchy

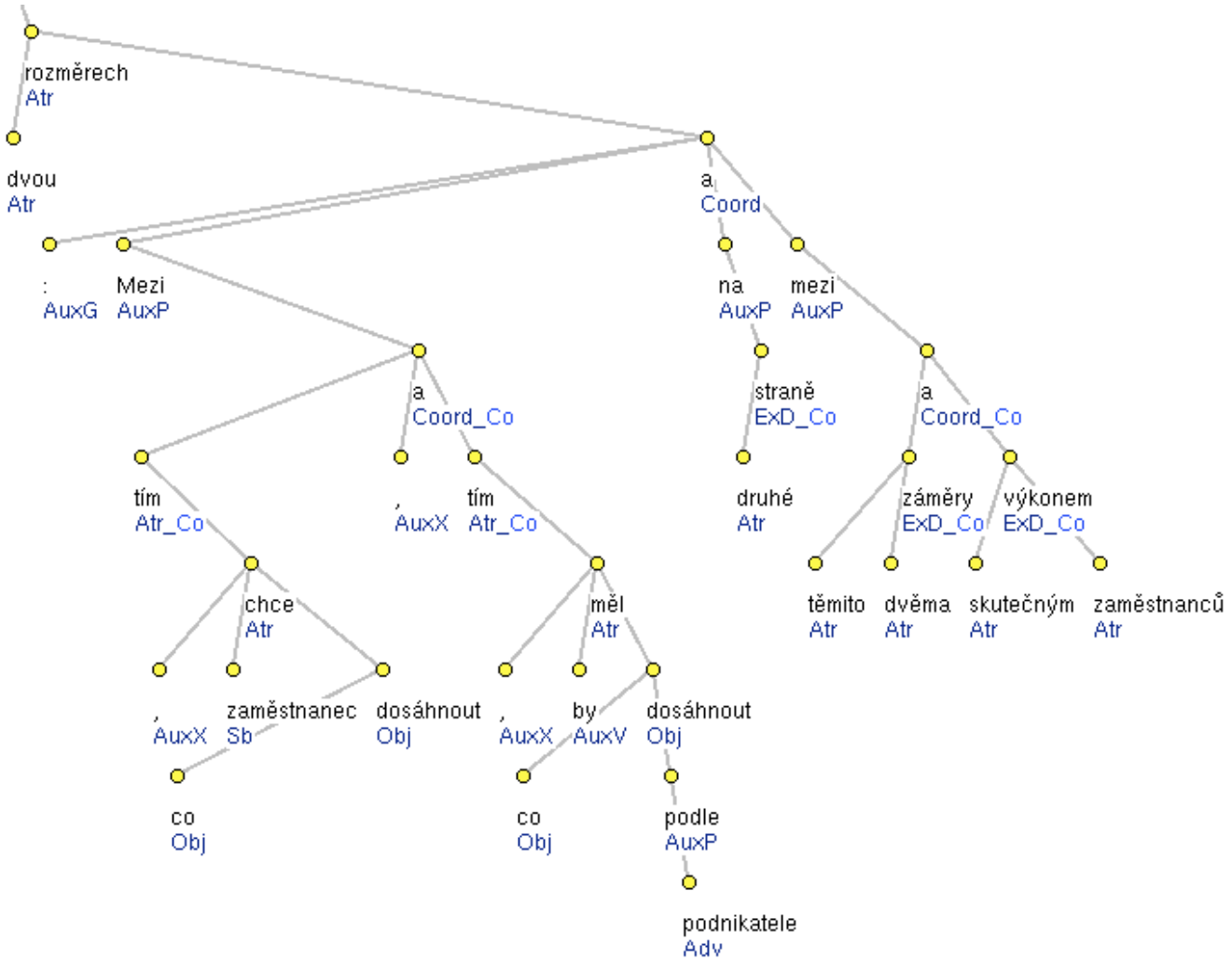
Remarkable Constructions II

- Ellipses
  - node(s) missing from the tree
- Prepositions and conjunctions
  - syntactic reasons, frequent exceptions
- ... these phenomena usually combine

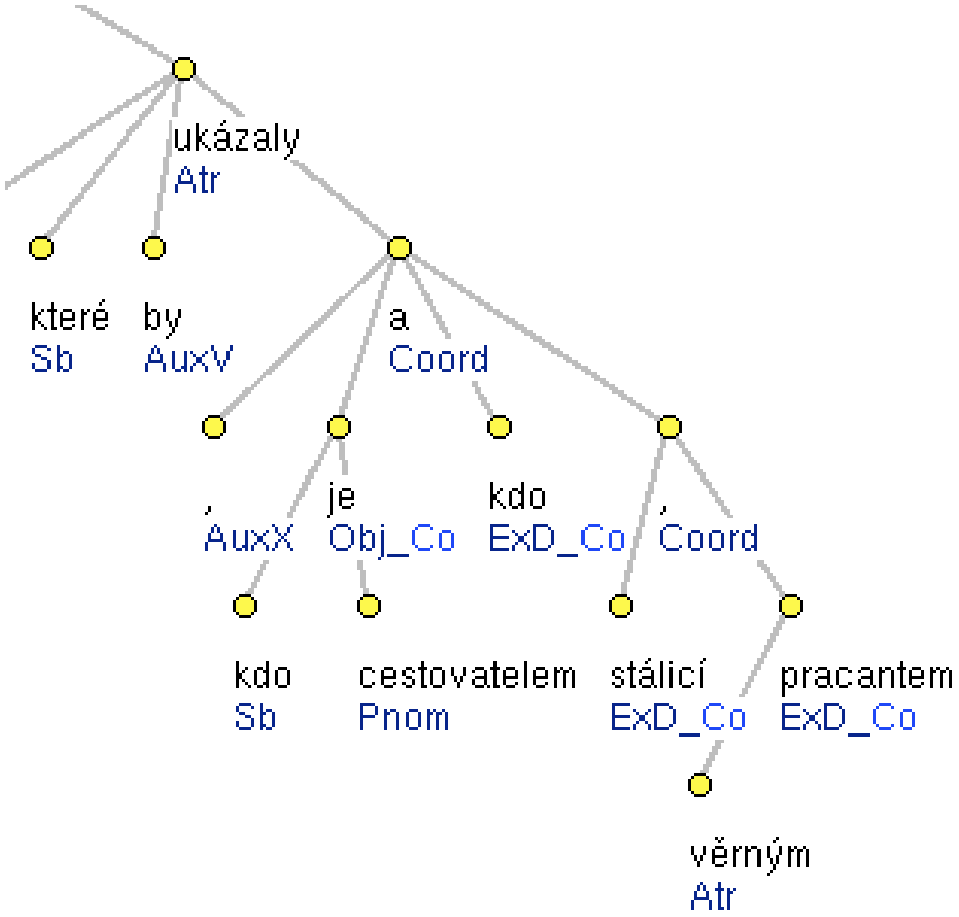
Consequences?



Consequences?



Consequences?



In Other Words . . .

- Traversing the tree - very non-trivial
- The `GetEChildren` function
- Uncertainty: Do I know everything?
- Influence by complex annotation rules
  I don't know (yet)?

PDT2vert

- Simplifying the corpus structure
  - linear structure is enough for usual NLP tasks
- The vertical format
- Structural tags

Structures

- Sentences
  - trivial
- Noun phrases
  - for each nominal node
- Clauses
  - very non-trivial
  - important for pruning NPs

Details

- The *btred* and the *fslib* library
  - in Perl
- `GetEChildren` + coordination analysis
- Linear changes
  - punctuation and symbols

Vertical file example

```
<sentence id="ln94202-55-p5s2">
<clause id="t-ln94202-55-p5s2w3">
<markable id="ln94202-55-p5s2w1" type="np" mtag="NNFS7-----A----">
Podmínkou    podmínka                NNFS7-----A----    Pnom
</markable>
však         však                    J^-------------    AuxY
je           být                     VB-S---3P-AA---    Pred
</clause>
,            ,                       Z:-------------    AuxX
<clause id="t-ln94202-55-p5s2w9">
aby          aby                     J,-------------    AuxC
<markable id="ln94202-55-p5s2w8" type="np" mtag="NNFP1-----A----">
tyto         tento                   PDFP1----------    Atr
činnosti     činnost_^(*3ý)          NNFP1-----A----    Sb
</markable>
s            s-1                     RR--7----------    AuxP
<markable id="ln94202-55-p5s2w12" type="np" mtag="NNIS7-----A----">
právním      právní                  AAIS7----1A----    Atr
úkonem       úkon                    NNIS7-----A----    Obj
obsaženým    obsažený_^(*5áhnout)    AAIS7----1A----    Atr
v            v-1                     RR--6----------    AuxP
<markable id="ln94202-55-p5s2w16" type="np" mtag="NNIS6-----A----">
notářském    notářský                AAIS6----1A----    Atr
zápise       zápis                   NNIS6-----A----    Adv
</markable>
</markable>
...
```

Further Work

- AR evaluation
- Experimenting with Information structure
- Evaluation of syntactic analysis
  - (phrase detection)
- anything needed :-)

Thank You for Your Attention!