An Annotated Corpus Outside Its Original Context

(A Corpus-Based Exercise Book)

Barbora Hladká, Ondřej Kučera
Institute of Formal and Applied Linguistics,
Charles University, Prague

Barbora Hladká, Ondřej Kučera: An Annotated Corpus Outside Its Original Context: A Corpus-Based Exercise Book
4th Workshop on Innovative Use of NLP for Building Educational Applications, ACL-08: HLT, Columbus, Ohio, 2008

Ultimate goal

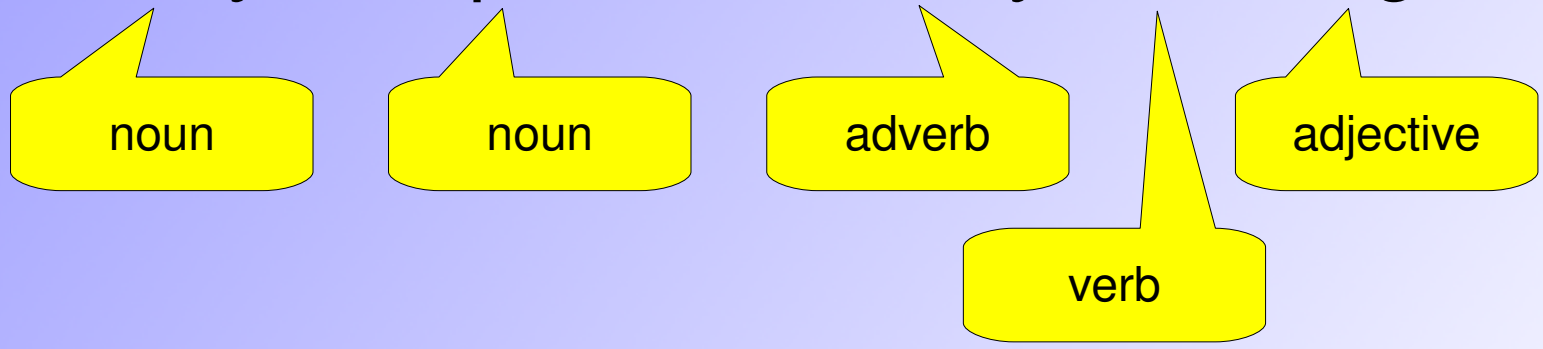
- students knowing (Czech) morphology and syntax

Ultimate goal (2)

- students knowing (Czech) morphology and syntax
- “Plat profesora opravdu není velký.”
  “A professor’s salary is not really high.”
  Lit.: “Salary of-a-professor really is-not high.”

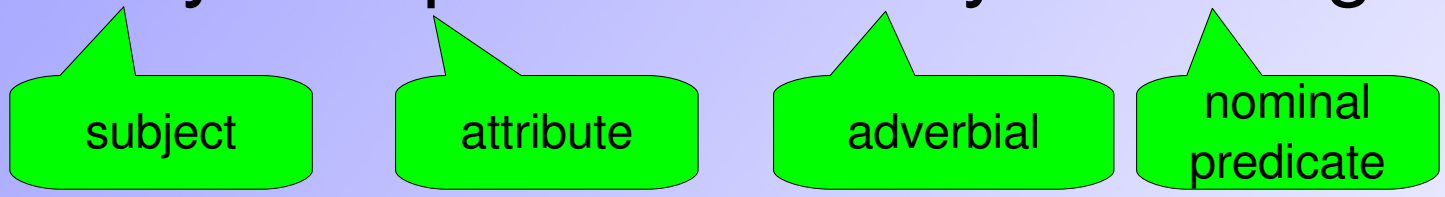
Ultimate goal (3)

- students knowing (Czech) **morphology** and syntax
- “Plat profesora opravdu není velký.”
  “A professor’s salary is not really high.”
  Lit.: “Salary of-a-professor really is-not high.”



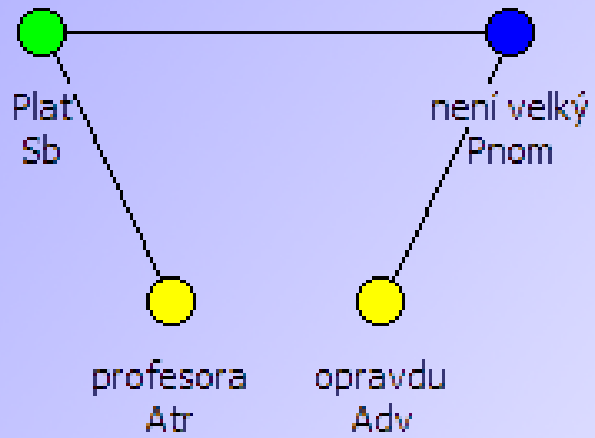
Ultimate goal (4)

- students knowing (Czech) morphology and **syntax**
- “Plat profesora opravdu není velký.”
  “A professor’s salary is not really high.”
  Lit.: “Salary of-a-professor really is-not high.”



Ultimate goal (5)

- students knowing (Czech) morphology and **syntax**
- “Plat profesora opravdu není velký.”
  “A professor’s salary is not really high.”
  Lit.: “Salary of-a-professor really is-not high.”



Getting there

- explaining
  - students have to be taught the rules
- understanding
  - then they have to understand it
- **exercising**
  - then they have to exercise it
  - this is what we're interested in

How to create an exercise book

- manually
  - author picks sentences from books, newspapers or makes them up
  - limited number of sentences
  - usually not very complicated sentences
  - time consuming (creating the key)
  - difficult not to make mistakes

How to create an exercise book (2)

- automatically
  - if we have an annotated corpus
  - number of exercises up to the volume of the corpus
  - “real life” sentences
  - hard work (annotating) already done
  - less errors

Our goal

- automatically built exercise book
  - because we have an annotated corpus
- complex parsing of a sentence
  - morphology
    - part of speech and other morphological categories (gender, number, case, …)
  - syntax
    - syntactic functions
    - graph of a sentence (dependency tree)

Prague Dependency Treebank

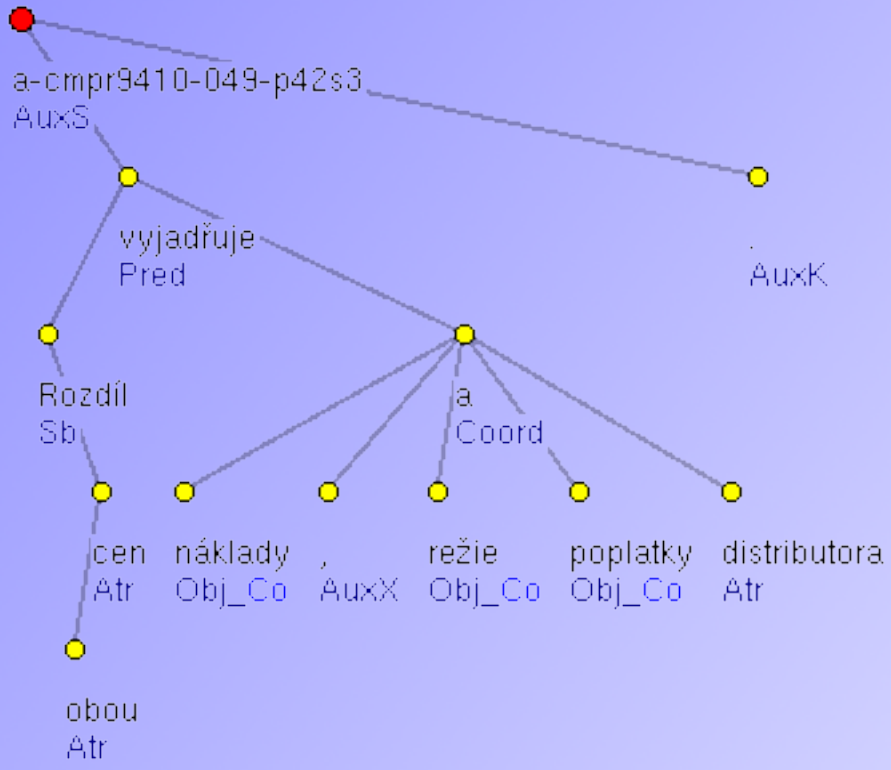
- Czech treebank of 2 million words
- three layers of annotation
  - morphological (2 mil. words)
  - syntactic (1.5 mil. words)
  - semantic (0.8 mil. words)
- http://ufal.mff.cuni.cz/pdt2.0/

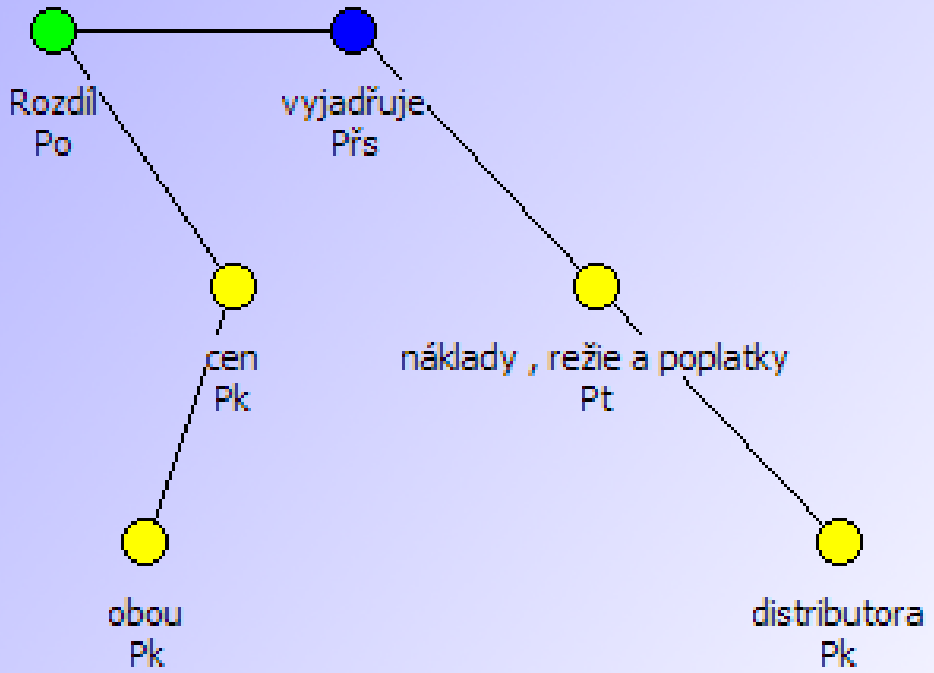
Processing PDT

- semantic layer: 49,442 sentences
- sentence filtering
  - “To často umožňuje, aby lidé pracovali na půl i méně plynu s tím, že když se to nebude šéfovi líbit, tak dotyčný půjde jinam - ke konkurenci.”
  - 11,705 sentences kept

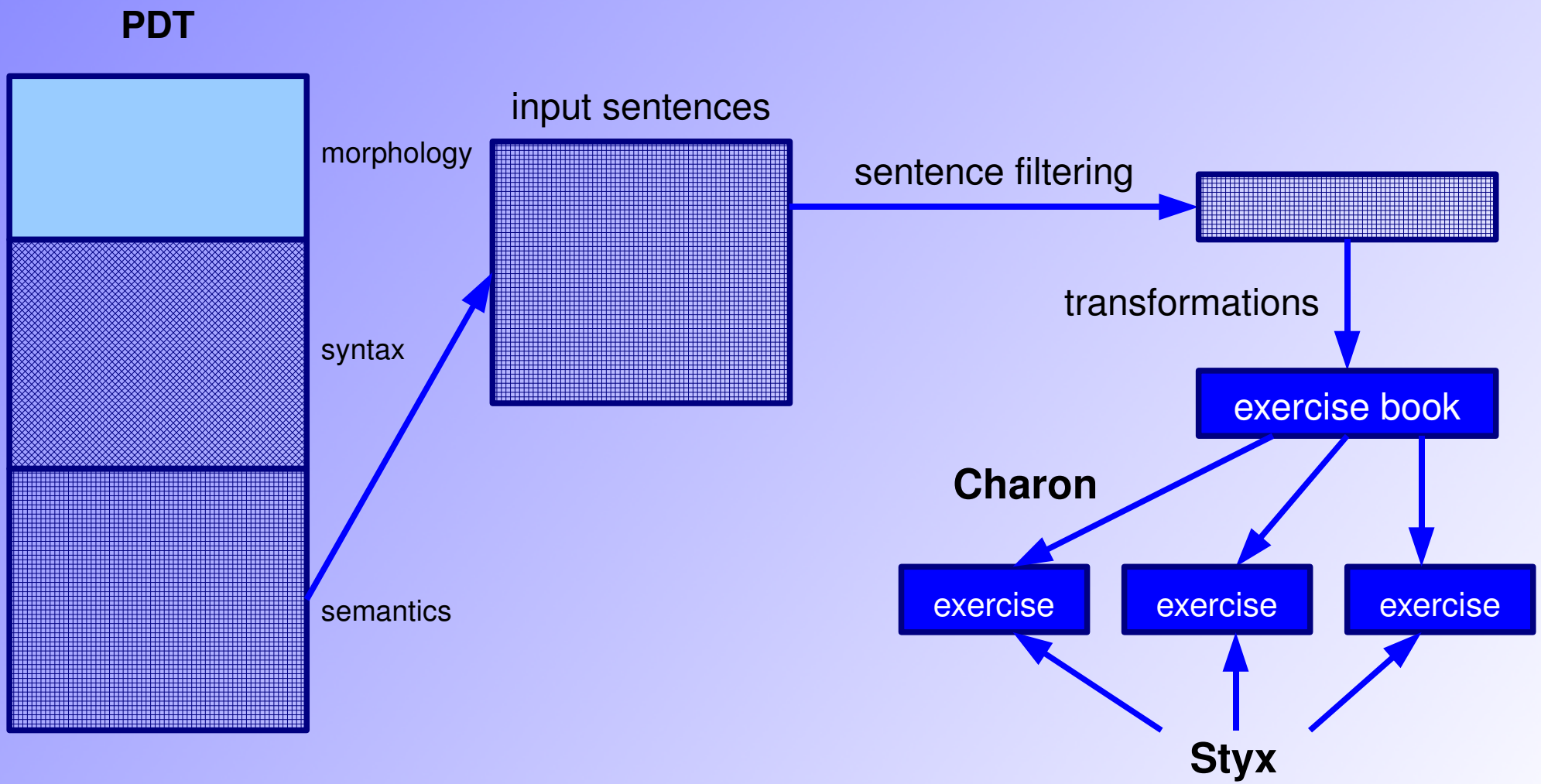
Processing PDT (2)

> syntactic transformations





Processing PDT (3)



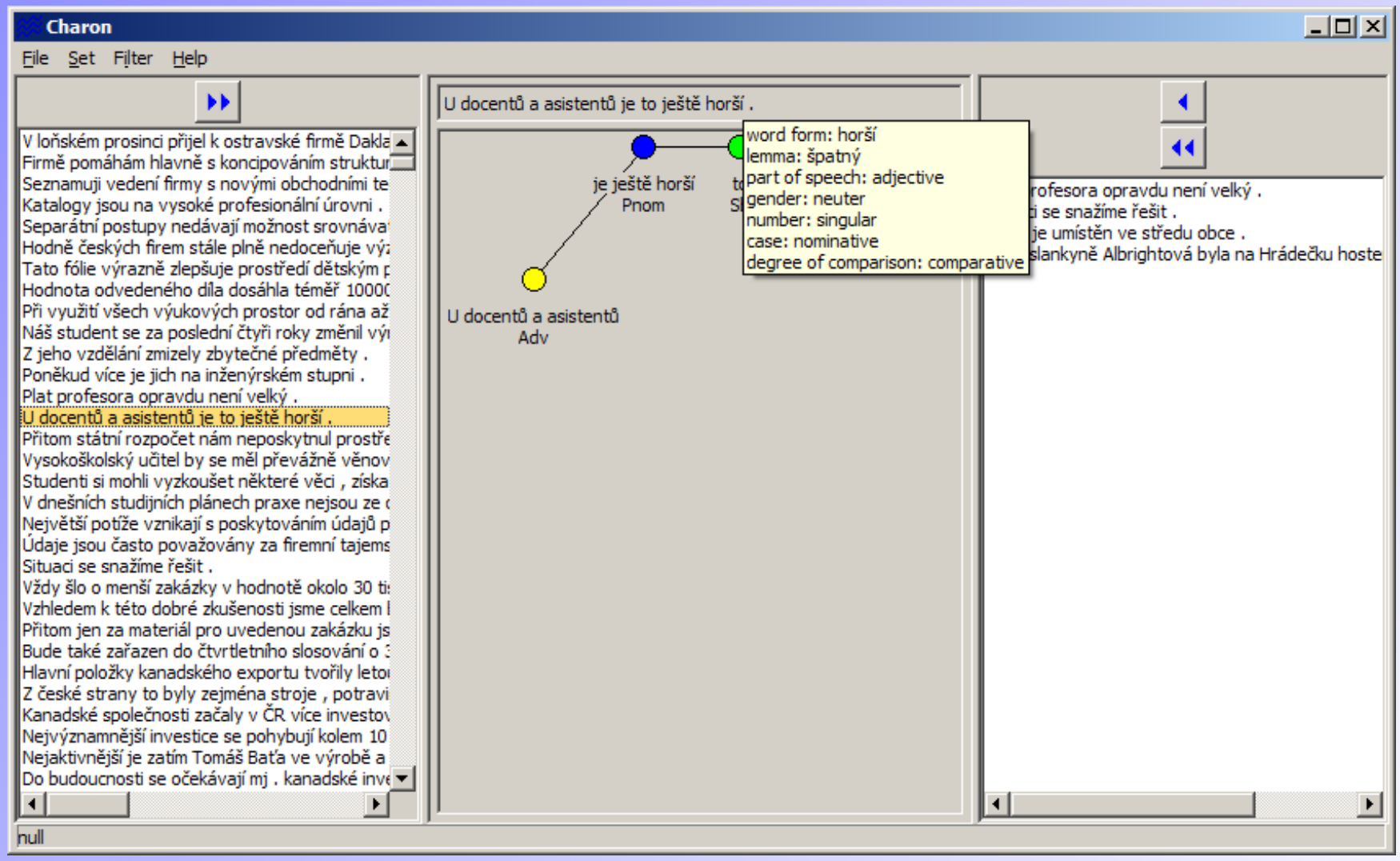
STYX

- automatically built electronic exercise book of Czech
- contains 11,705 sentences
- three applications
  - FilterSentences
  - Charon
  - Styx

Charon

- intended primarily for teachers
- user sees all sentences in the exercise book
  - the view can be filtered
    - by presence/absence of some phenomena
    - show only sentences containing verbal predicate and not containing any adverbials or attributes
- user selects some sentences and creates an exercise from them

Charon (2)

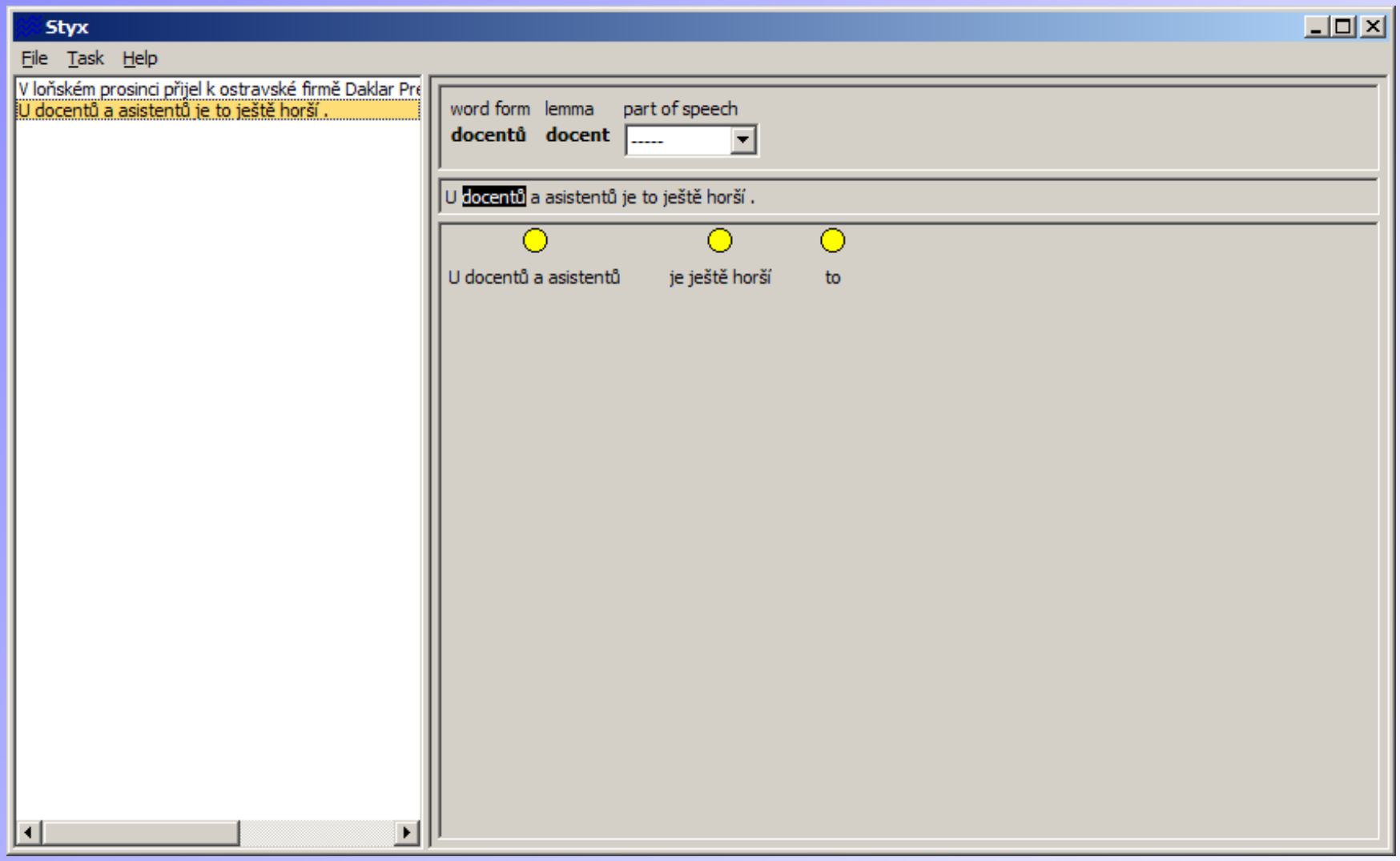


Barbora Hladká, Ondřej Kučera: A Corpus-Based Exercise Book
ACL-08: HLT, Columbus, Ohio, 2008

Styx

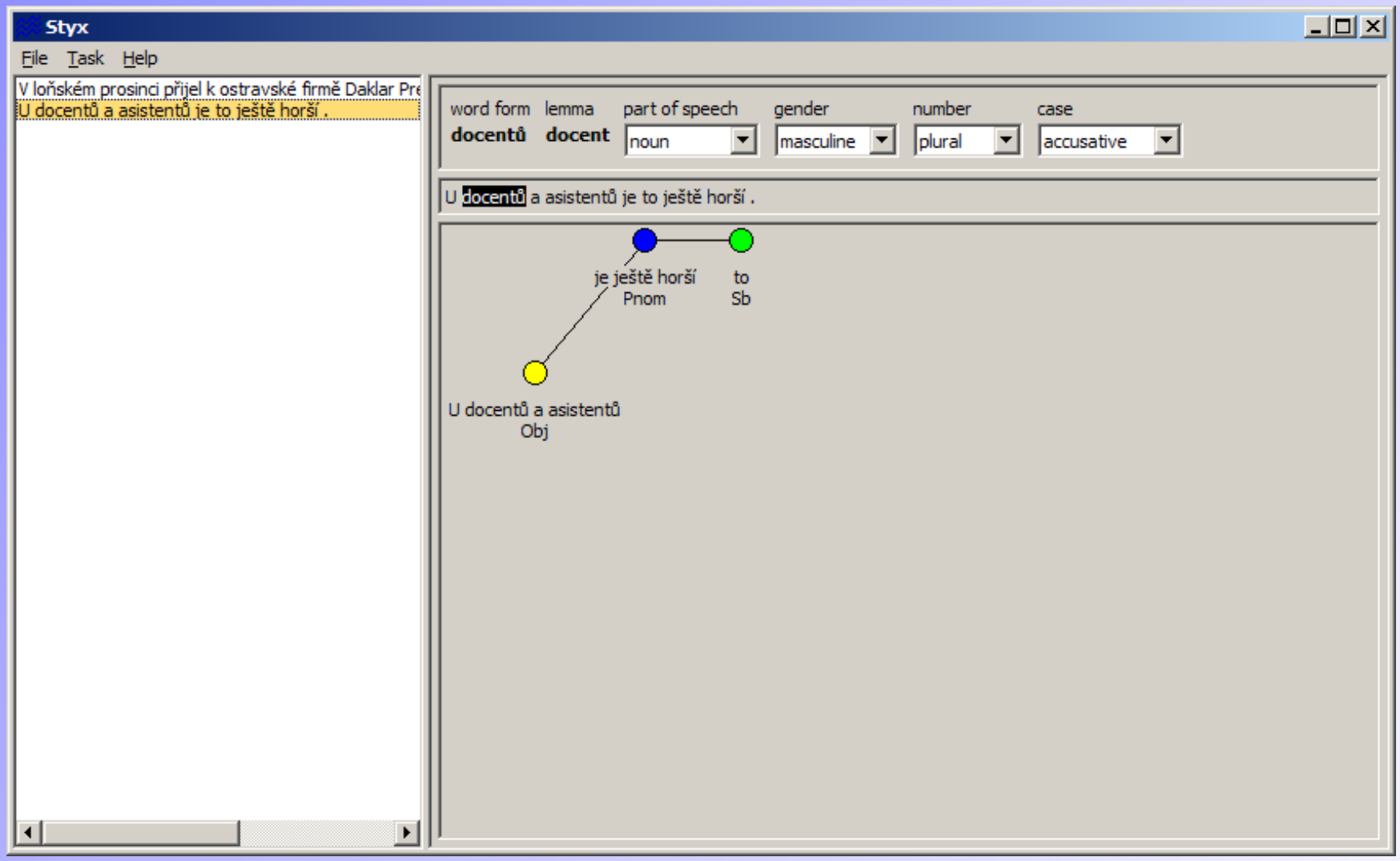
- exercise book itself
- user loads an exercise created with Charon
- … and practices
- in the end users can check their results against correct solutions

Styx (2)



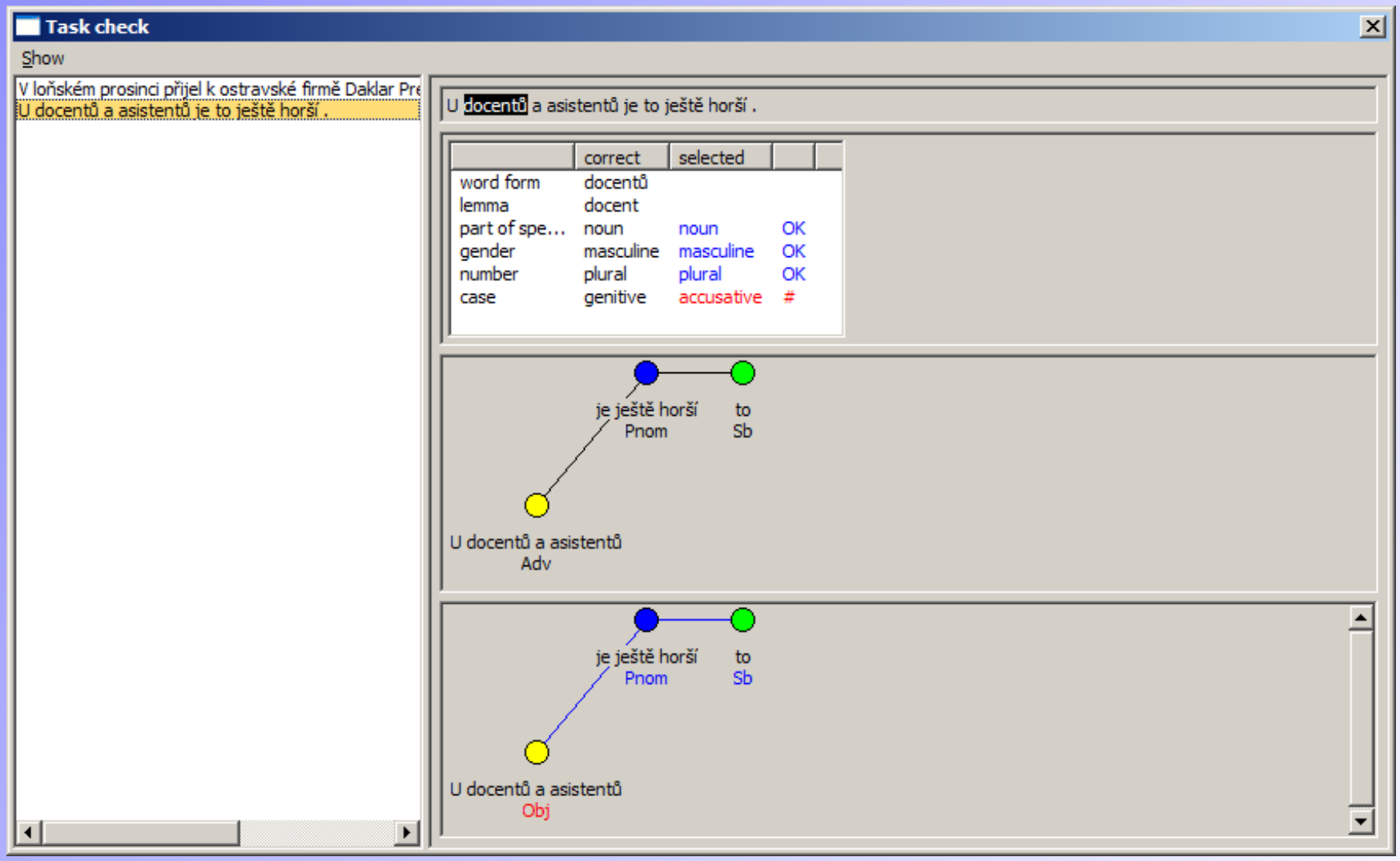
Barbora Hladká, Ondřej Kučera: A Corpus-Based Exercise Book
ACL-08: HLT, Columbus, Ohio, 2008

Styx (3)



Barbora Hladká, Ondřej Kučera: A Corpus-Based Exercise Book
ACL-08: HLT, Columbus, Ohio, 2008

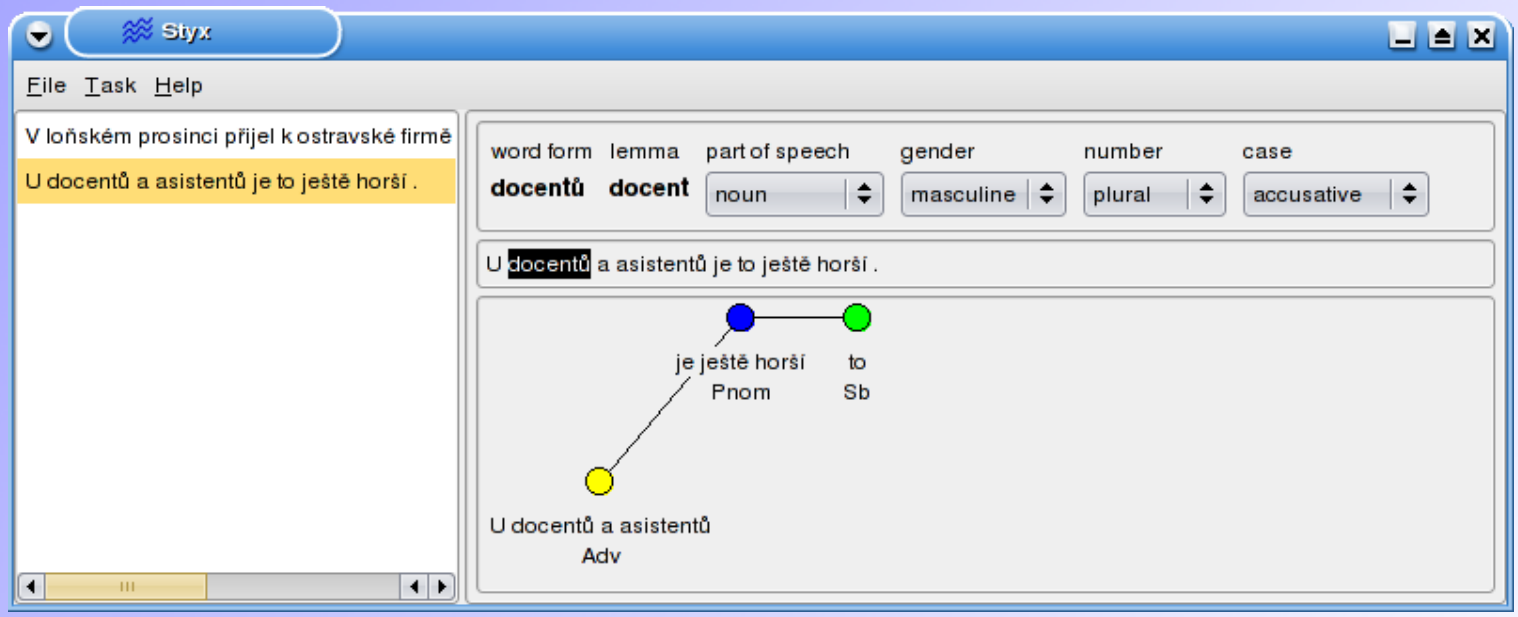
Styx (4)



Barbora Hladká, Ondřej Kučera: A Corpus-Based Exercise Book
ACL-08: HLT, Columbus, Ohio, 2008

Implementation

- Java
- SWT (Standard Widget Toolkit)
  - native look and feel on each platform
- GPL



Present and future

- current version 0.9.2
- language improvements
  - kinds of attributes (concordant, discordant)
  - kinds of adverbials (time, place, manner, …)
- user interface improvements

http://ufal.mff.cuni.cz/styx/

Barbora Hladká, Ondřej Kučera: A Corpus-Based Exercise Book
ACL-08: HLT, Columbus, Ohio, 2008