

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

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.”

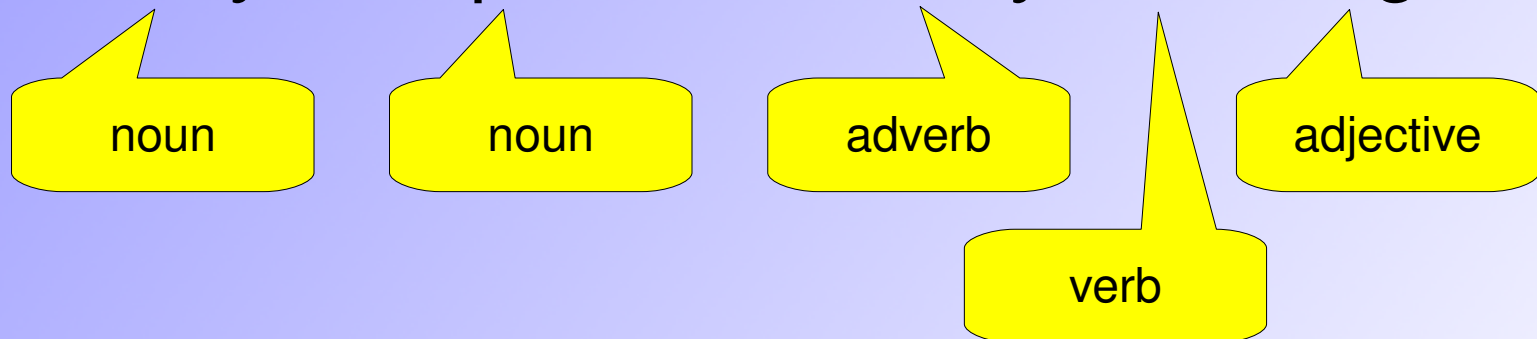
Ultimate goal (3)

> 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.”

subject

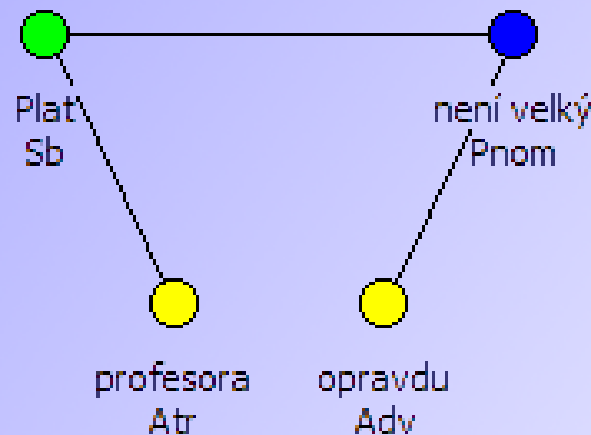
attribute

adverbial

nominal
predicate

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

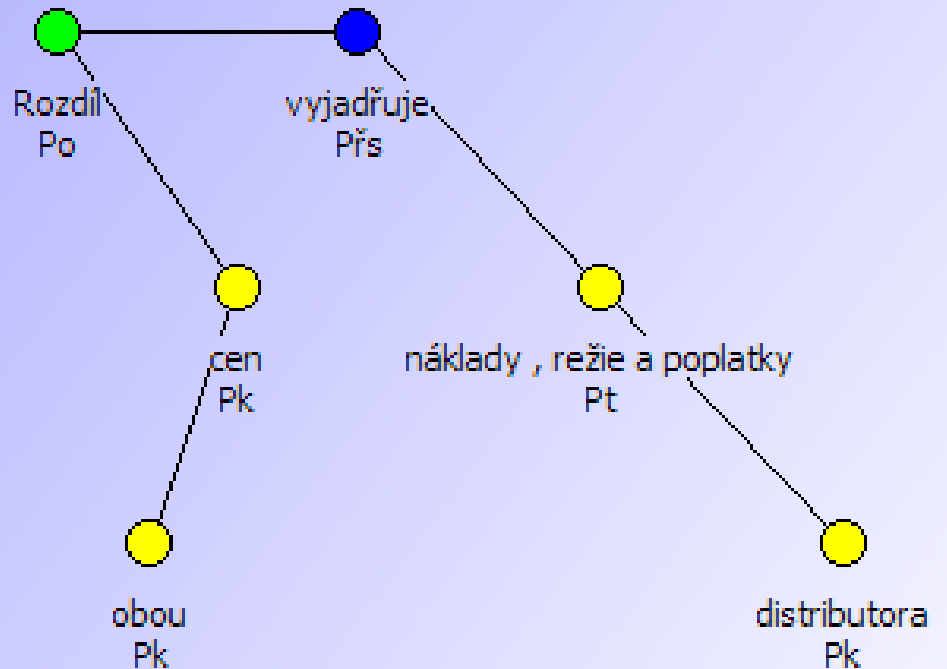
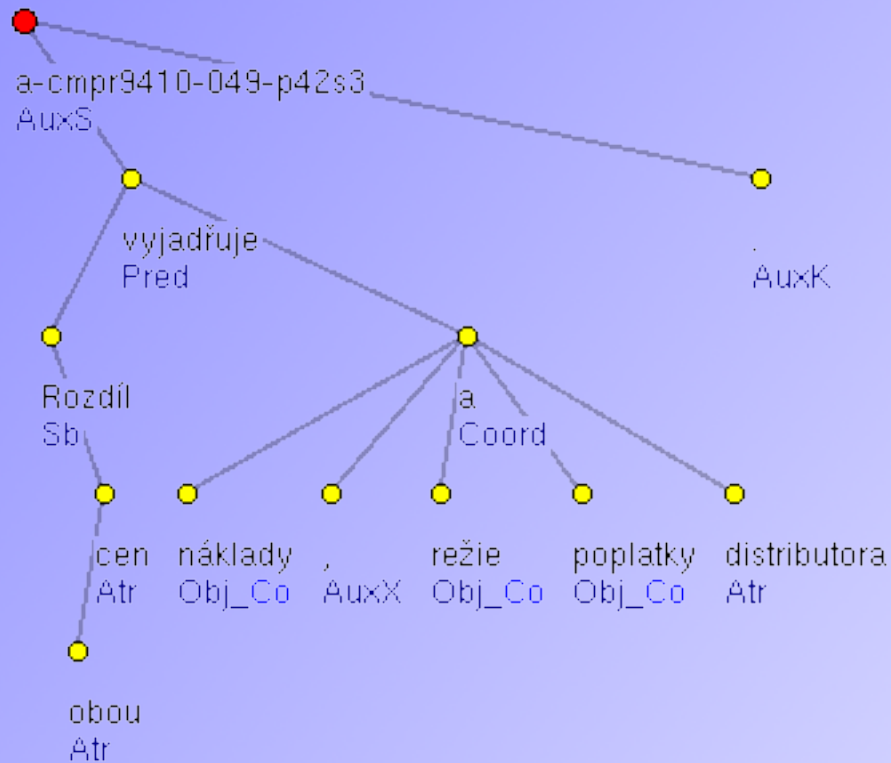
- > Czech treebank of 2 million words
- > three layers of annotation
 - > morphological (2 mil. words)
 - > syntactic (1.5 mil. words)
 - > semantic (0.8 mil. words)
- > <http://ufal.mff.cuni.cz/pdt2.0/>

Processing PDT

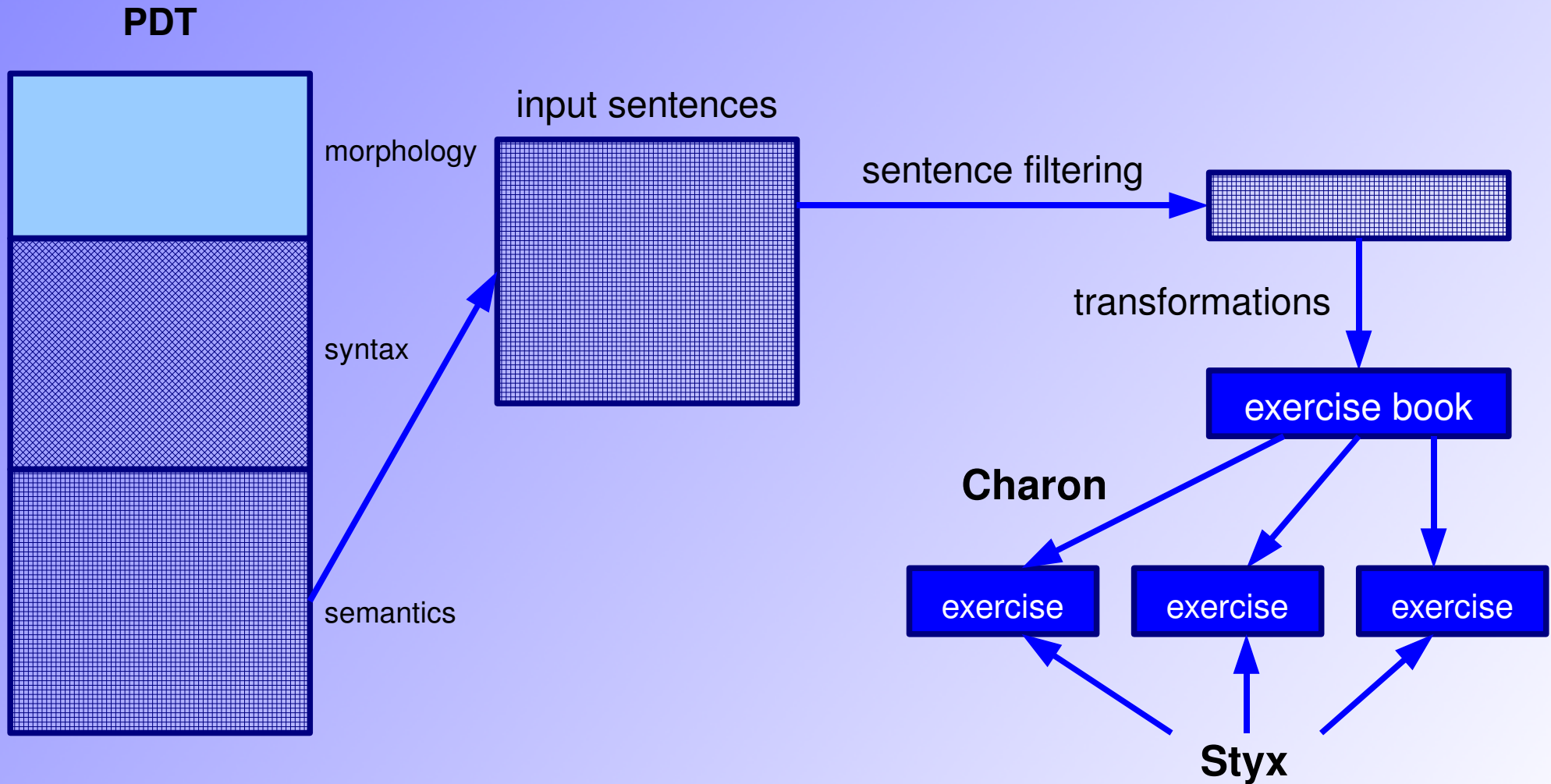
- > semantic layer: 49,442 sentences
- > sentence filtering
 - > “To často umožňuje, aby lidé pracovali na půl i méně plynu s tím, že když se to nebude šéfovi líbit, tak dotyčný půjde jinam - ke konkurenci.”
 - > 11,705 sentences kept

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)

The screenshot shows the Charon software interface. The main window displays a sentence: "U docentů a asistentů je to ještě horší ." The word "horší" is highlighted in yellow. A tooltip window is open over the word, displaying the following morphological information:

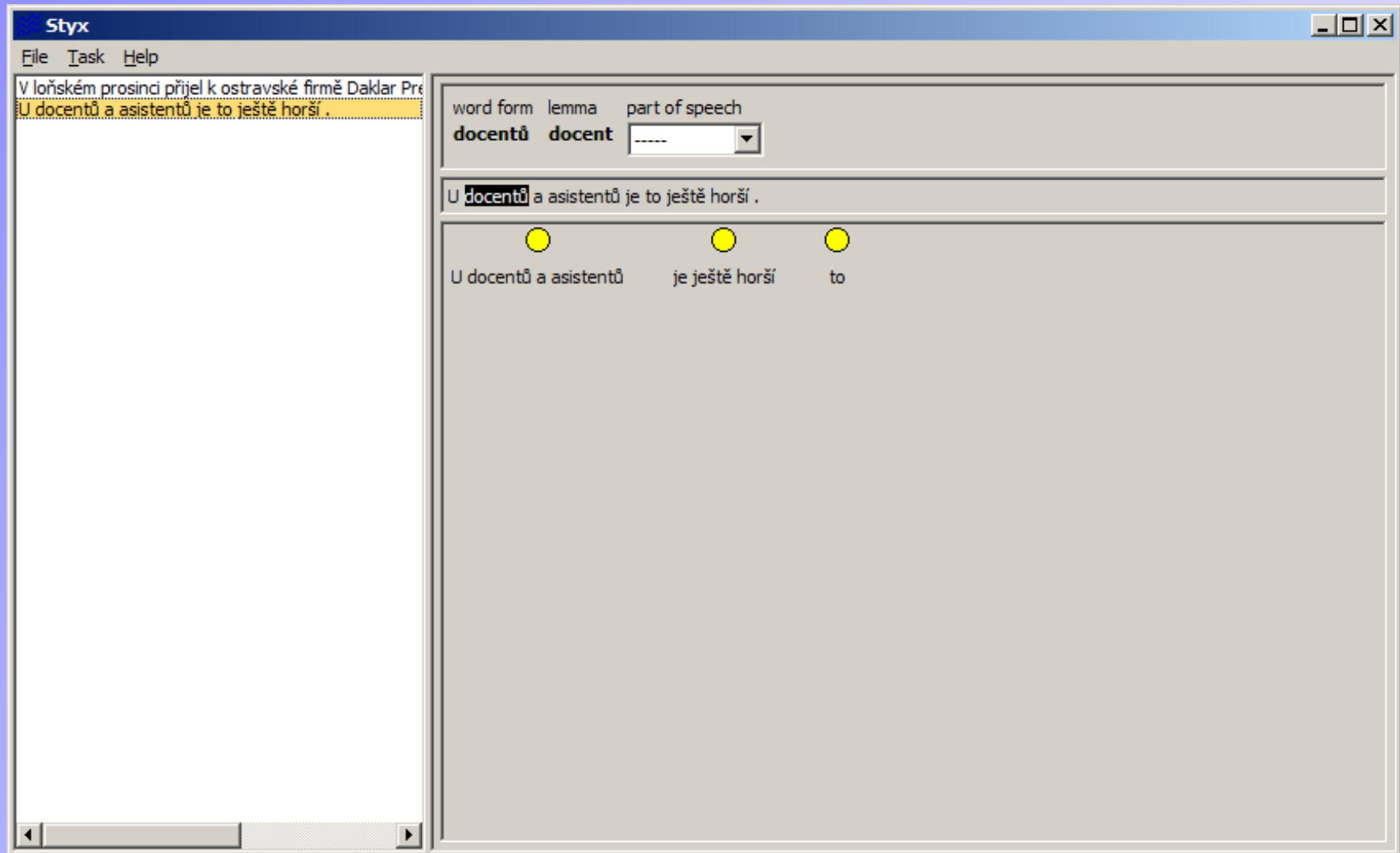
- word form: horší
- lemma: špatný
- part of speech: adjective
- gender: neuter
- number: singular
- case: nominative
- degree of comparison: comparative

The interface also shows a list of words on the left side, including "je ještě horší" (Pnom) and "U docentů a asistentů" (Adv). The bottom of the window shows the word "null".

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)



Styx (3)

The screenshot shows the Styx software interface. The window title is "Styx". The menu bar includes "File", "Task", and "Help". The main text area contains the sentence "U loňském prosinci přijel k ostravské firmě Daklar Pre" and the highlighted phrase "U docentů a asistentů je to ještě horší .".

Below the text area, there is a morphological analysis table:

word form	lemma	part of speech	gender	number	case
docentů	docent	noun	masculine	plural	accusative

Below the table, the sentence "U **docentů** a asistentů je to ještě horší ." is shown. A diagram illustrates the syntactic structure:

- A yellow circle represents the object "U docentů a asistentů" (Obj).
- A blue circle represents the subject "je ještě horší" (Pnom).
- A green circle represents the object "to" (Sb).
- Lines connect the yellow circle to the blue circle, and the blue circle to the green circle.

Styx (4)

Task check

Show

V loňském prosinci přijel k ostravské firmě Daklar Pre
U docentů a asistentů je to ještě horší .

U **docentů** a asistentů je to ještě horší .

	correct	selected		
word form	docentů			
lemma	docent			
part of spe...	noun	noun	OK	
gender	masculine	masculine	OK	
number	plural	plural	OK	
case	genitive	accusative	#	

je ještě horší to

Pnom Sb

U docentů a asistentů

Adv

je ještě horší to

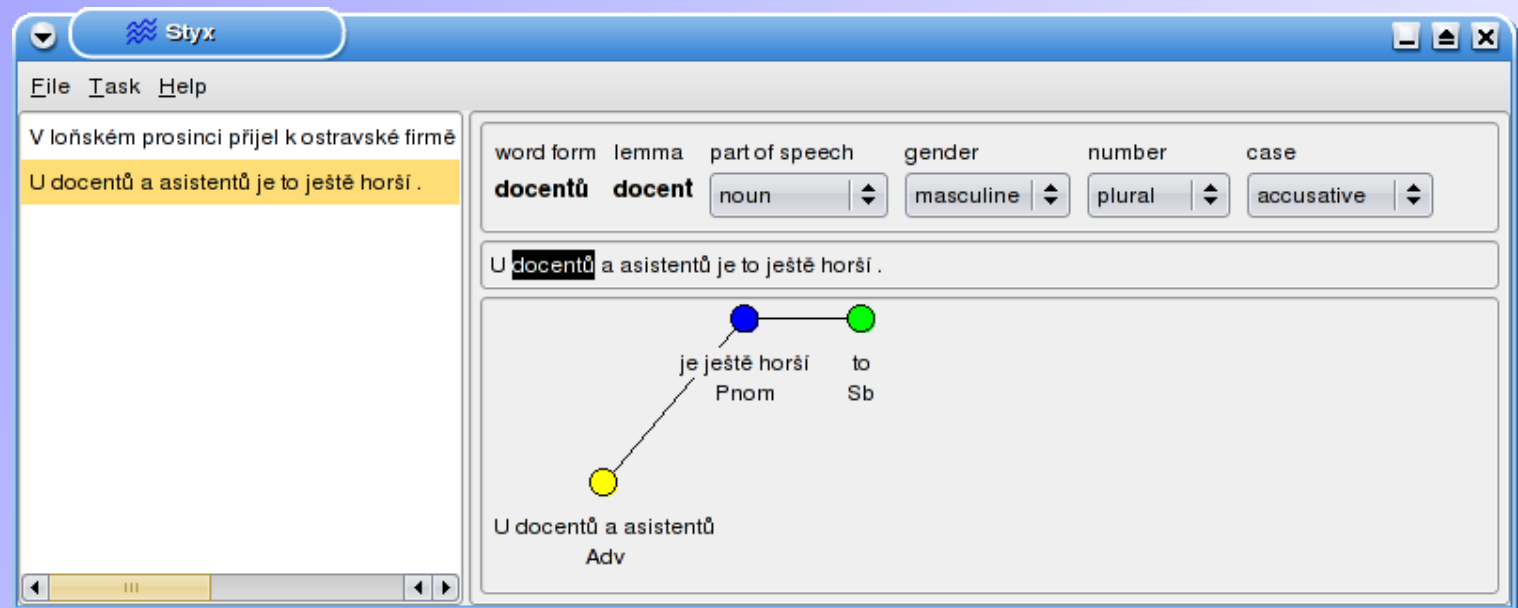
Pnom Sb

U docentů a asistentů

Obj

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/



STYX Prague Dependency Treebank as an Exercise Book of Czech Prague Dependency Treebank as an Exercise Book of Czech Prague Dependency Tree
Pražský závislostní korpus jako cvičebnice češtiny Pražský závislostní
ský závislostní korpus jako cvičebnice češtiny Pražský závislostní korpus

[MORE DETAILS](#) Play the Language [PODROBNĚJŠÍ INFORMACE](#) Hraj si se slovy a s větami

The **STYX** is a system designed to provide an **electronic corpus-based exercise book of Czech** morphology and syntax with exercises directly selected from the **Prague Dependency Treebank**, the largest annotated corpus of Czech ([see sample data](#)).

■ Exercises give **practice in**

- classifying part of speech and particular morphological categories (such as gender, number, case, tense, ...) of words
- parsing a sentence and classifying syntactic function (such as subject, predicate, objects, ...) of words

■ **Correct answers** at hand

■ **Java** implementation

■ **Publications, Presentations & Awards**

- For more information click: [here](#)

■ **Download leaflet** ([pdf](#))

Systém **STYX** je **elektronickou cvičebnicí českého tvarosloví a syntaxe** sestavenou na základě anotovaného korpusu. Cvičení jsou vybraná z **Pražského závislostního korpusu**, největšího anotovaného korpusu českých textů ([viz ukázka](#)).

■ Ve cvičeních **se procvičuje**

- určování slovních druhů a příslušných morfologických kategorií slov věty (rod, číslo, pád, čas ...)
- dělání větného rozboru a určování větných členů slov věty (podmět, přísudek, předmět, ...)

■ **Klíč k řešení**

■ Implementace v prostředí **Java**

■ **Publikace, Presentace & Ocenění**

- Podrobný výčet je uveden [zde](#)

■ **Plakát ke stažení** ([pdf](#))

[CONTACT](#) [Barbora Hladká](#), [Ondřej Kučera](#)

Show Images 100%