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.”
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
> 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)
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)
Styx (4)
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/

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 vyznačování a syntaxe sestavenou na základě anotovaného korpusu. Čvíčení jsou vybraná z Právěské závislostního korpusu, největšího anotovaného korpusu českých textů (viz uladka).

- Ve cvičení se provádí
  - určování slovních druhů a příslušných morfológych kategorií slov věty (rod, číslo, pád, čas, ...)
  - určování větového rozeznávání a určování větových druhů slov věty (podmínka, případ, předmět, ...)

- Křížení
  - Implementace v prostředí Java
  - Publikace, Prezentace & Ocenění
    - Podrobný výchoj je uveden zde
    - Plakát ke stažení (pdf)

CONTACT Barbora Hladká, Ondřej Kučera