# Víceúrovňová anotace českého žákovského korpusu

Svatava Škodová[1] Barbora Štindlová[1]
Jirka Hana[2] Alexandr Rosen[2]

[1]Technická Univerzita, Liberec
[2]Karlova univerzita, Praha

Korpling 2011
Korpusová lingvistika
Praha, 22.–24. září 2011

# Obsah

# Obsah

# Žákovský korpus

- elektronický soubor jazykových projevů studentů daného jazyka jako jazyka cizího/druhého (Leech 1998)
- 1990 – ICLE (S. Granger, Université catholique de Louvain)

# CzeSL – žákovský korpus češtiny

- první velký žákovský korpus pro slovanský jazyk
  – slovinština: PiKUST (35 tis. slov)
- první korpus češtiny jako druhého / cizího jazyka
- součást projektu *AKCES* – akviziční korpusy češtiny

# Plánovaný rozsah v roce 2012

- 1 mil. slov
- 3 subkorpusy podle prvního jazyka
- psaná a mluvená část
- všechny úrovně znalosti jazyka podle SERR

# Obsah

| Korpus | Mil. slov | L1 | L2 | Úroveň | Médium | Anotace |
|---|---|---|---|---|---|---|
| ICLE | 3 | 26 | en | pokr. | psaný | část |
| CLC | 35 | 130 | en | všechny | psaný | část |
| LINDSEI | 0,8 | 11 | en | pokr. | mluv. | část |
| PELCRA | 0,5 | pl | en | všechny | psaný | část |
| USE | 1,2 | sv | en | pokr. | psaný | ne |
| HKUST | 25 | zh | en | pokr. | psaný | část |
| CHUNGDAHM | 131 | ko | en | všechny | psaný | část |
| JEFLL | 0,7 | jp | en | zač. | psaný | část |
| MELD | 1 | 16 | en | pokr. | psaný | ne |
| MICASE | 1,8 | různé | en | pokr. | mluv. | ne |
| NICT JLE | 2 | jp | en | všechny | mluv. | část |
| FALKO | 0,3 | 5 | de | pokr. | psaný | část |
| FRIDA | 0,2 | různé | fr | stř.pokr. | mluv. | část |
| FLLOC | 2 | en | fr | všechny | mluv. | ne |
| PiKUST | 0,04 | 18 | sl | pokr. | psaný | ano |
| ASU | 0,5 | různé | no | pokr. | psaný | ne |
| TUFS | znaků: 0,6 | různé | jp | všechny | psaný | ne |

# Anotace žákovských korpusů

Data nerodilých mluvčích se v žákovských korpusech mohou anotovat dvěma na sobě nezávislými způsoby:

## Lingvistické značkování

- lemmatizace
- slovnědruhová, morfologická, příp. syntaktická anotace
- využití softwarových nástrojů pro analýzu národního jazyka

## Chybová anotace

- manuální
- úroveň, rozsah a koncept chybových anotací se značně odlišují

# Chybová anotace

- přibližně 46 % žákovských korpusů je anotovaných
- selektivní chybová anotace:
  - výslovnost (LeaP)
  - pravopis (TLEC)
  - syntax (AleSKO)
- komplexní chybová anotace:
  ICLE, FRIDA. FALCO, NICT JLE, CzeSL

# Možnosti zachycení chyb

## 1. **Rekonstrukce** – implicitní zachycení chyb

- chyba v textu je detekována a nahrazena korektní formou
- absence klasifikačního schématu (Fitzpatrick a Seegmiller 2004)
  - výhody
    - anotátor se ho nemusí učit
    - anotování je rychlejší
    - nedochází k chybnému zařazení, příp. přehlížení chyb
  - nevýhody
    - omezený počet anotátorů
    - problematické pro automatickou analýzu
    - problemtické pro uživatele

## 2. **Klasifikace** – explicitní zachycení chyb

- chybová taxonomie
  - identifikace a kategorizace chyb podle předem vymezené chybové typologie
- vždy odráží teoretický koncept, v jehož rámci vznikla

# Typy chybových taxonomií

## Lingvisticky zaměřené taxonomie

- různě podrobné
  - od široce pojatých kategorií (morfologie, lexikum, syntax) ke kategoriím specifickým (pomocná slovesa, pasivum, apod.)
  - hierarchicky uspořádané,
    - chybová doména (gramatická, lexikální, stylová)
    - chybová kategorie (např. diakritika, flexe, rod, modus, atp.)
    - slovní druh (POS)

## Taxonomie podle povrchové realizace

- formální typy alternace zdrojového textu (chybějící element, přebývající element, chybně utvořený element, chybné uspořádání)
- často jako komplementární k lingvisticky orientované kategorizaci

# Obsah

# Chybová anotace v korpusu CzeSL

- problematické jevy v češtině
  - flexe, derivace, shoda, slovosled podle aktuálního členění ap.

## Řešení

- víceúrovňové anotační schéma
- kombinace manuální a automatické anotace

## Automatická anotace

- automatické přiřazování chybových značek na základě porovnání chybných a opravených tvarů
- morfosyntaktické značkování a lemmatizace

Viktor je mladý pan z Ruska. Studuje češtinu ve škole, protože ne umí psat a číst správně. Bydlí na kolejе vedle školy, má jednu sestru Irenu, která se učí na univerzitě u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra píše všechno a výborně rozumí českému profesora Smutneveselého a brzo dělá domácí úkol. Večeře Irena jde na prohaska spolu z kamarádem, ale její bratr dělá nic. Jelo čeština je špatná, vím, že se vrátit ve Rusku Polsko a tam bude studovat a pomalu myt podlahy.

Kamarad Ireny je Američan a chytrý muž. On miluje Irenu a chce se vzít na ní, protože ona je hezká, taky chytrá, rozumí ho a umí výborně

<li>Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrym student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveselého {a brzo delá domací ukol}<in>. Večeře Irena jde na prohaska spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Rusk~~o~~u a tam budí studovat u pomalu myt podlahy.

<li>Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborný vařit.

<li>Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a číst spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrym student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveselého {a brzo delá domací ukol}<in>. Večeře Irena jde na prohaska spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Rusk~~o~~u a tam budí studovat u pomalu myt podlahy.

<li>Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborný vařit.

<li>Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrym student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveselého {a brzo delá domací ukol}<in>. Večeře Irena jde na prohaska spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Ruskou a tam budí studovat u pomalu myt podlahy.

<li>Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborný vařit.

# Anotační schéma

## Třírovinný formát a dvoustupňová anotace

umožňuje

- postupnou emendaci
- anotaci chyb v izolovaných tvarech i (nespojitých) řetězcích

# Roviny anotace

## ROVINA 0

- přepis původního textu

## ROVINA 1

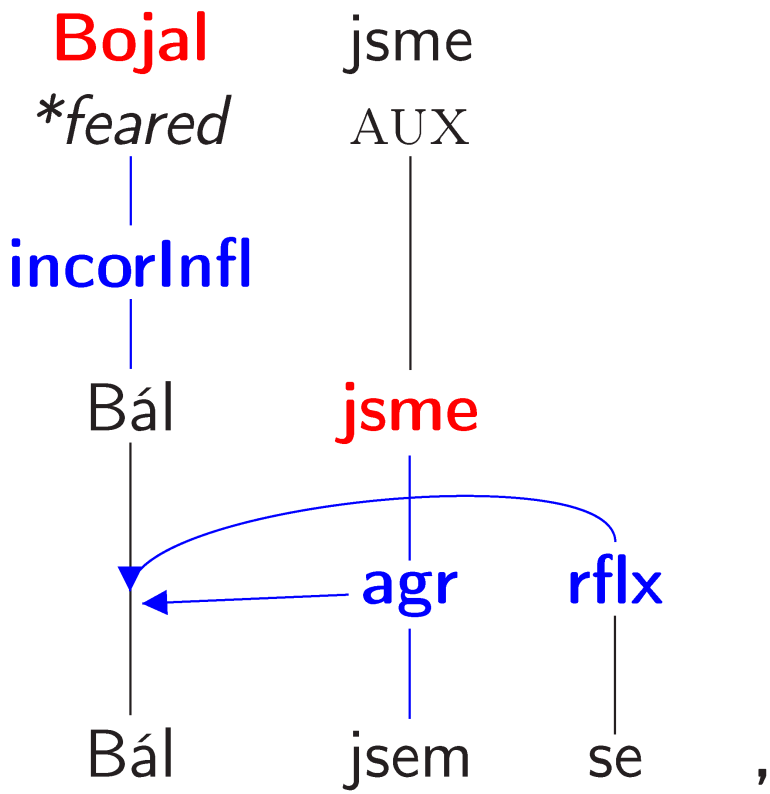
- oprava pravopisných a morfologických chyb izolovaných tvarů
- výsledek:
  - řetězec existujících českých forem
  - věta jako celek může být chybná

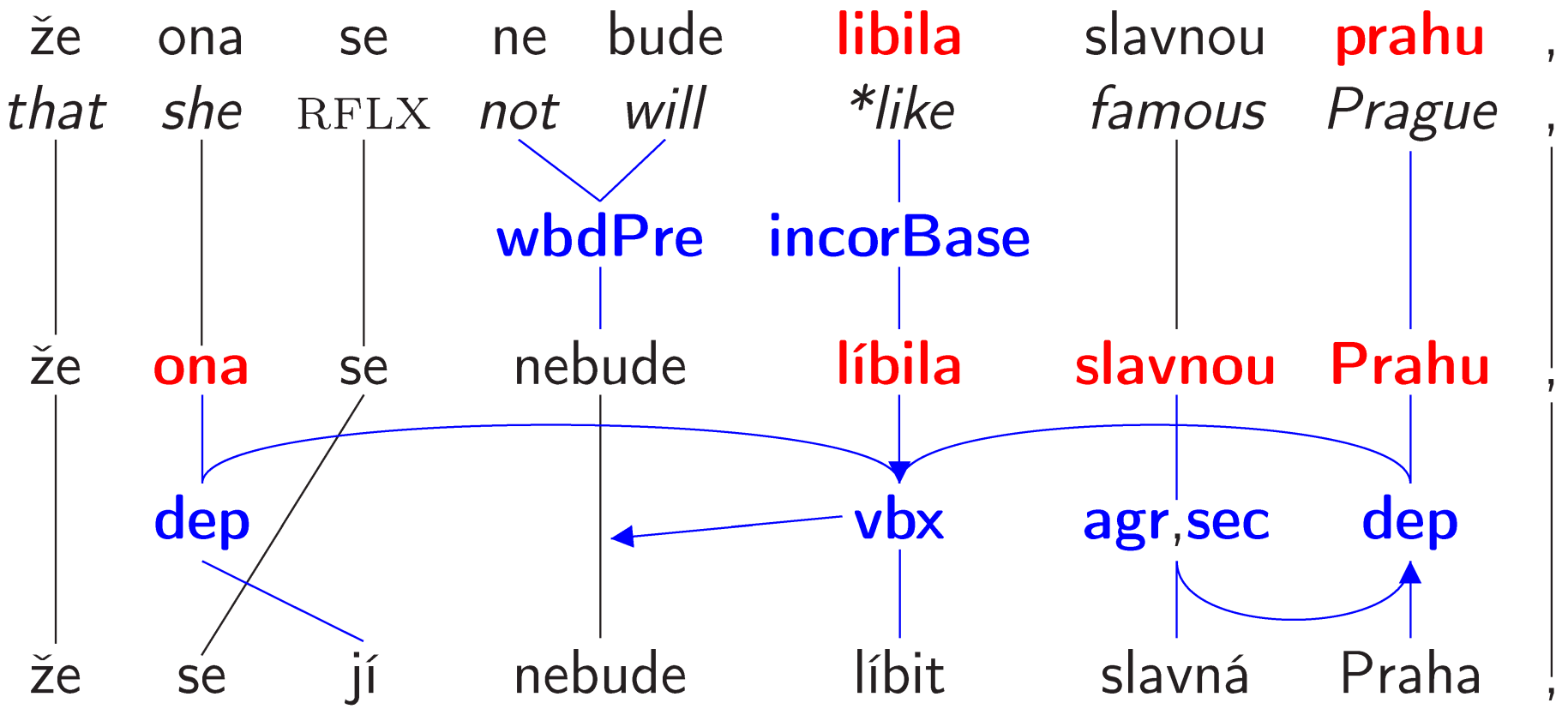
## ROVINA 2

- ostatní typy chyb (syntaktické, lexikální, slovosledné, stylové, v referenci, negaci atd.)
- výsledek:
  - gramaticky správná věta

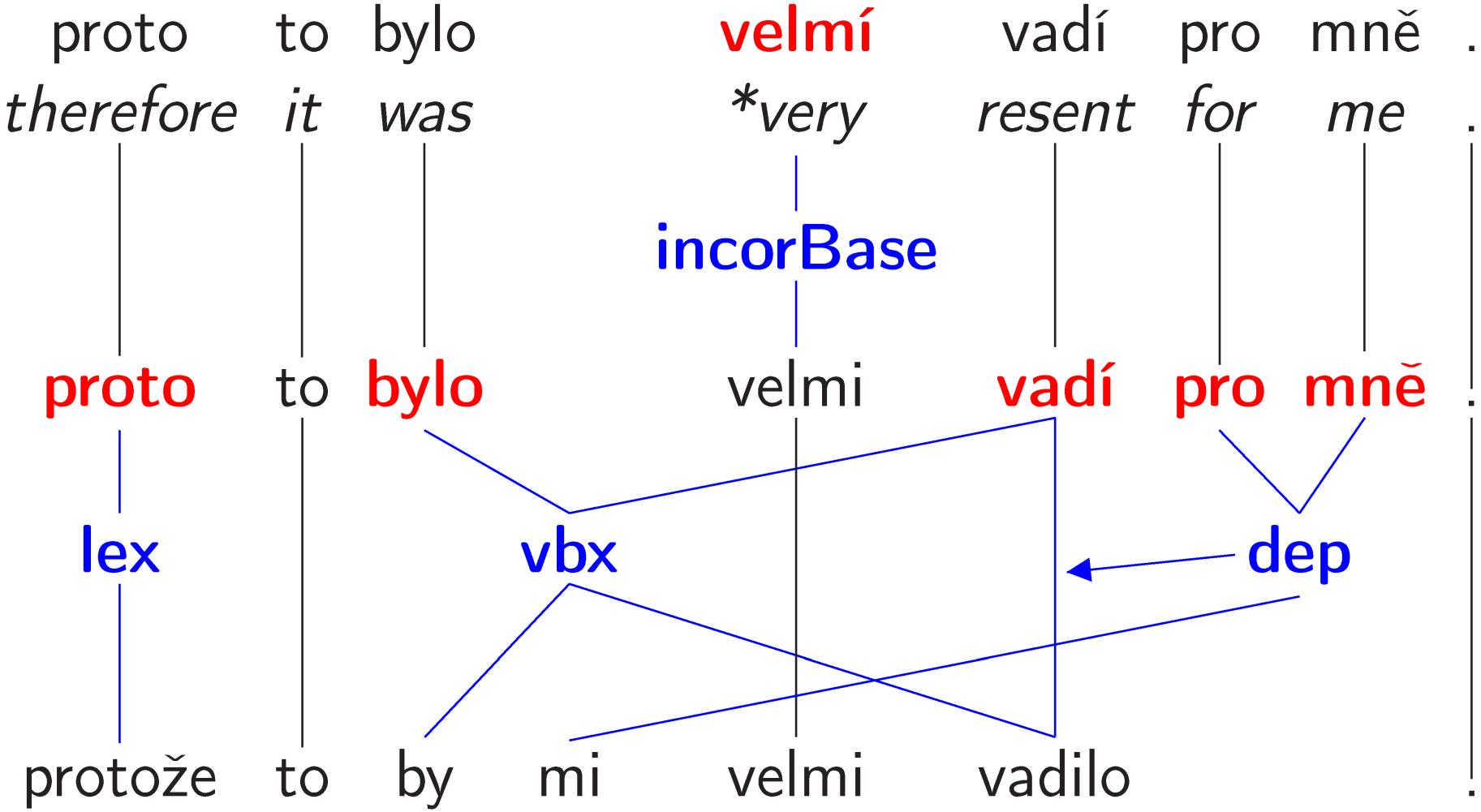
**Bojal jsme se že ona se ne bude libila slavnou prahu,**

**proto to bylo velmí vadí pro mně.**

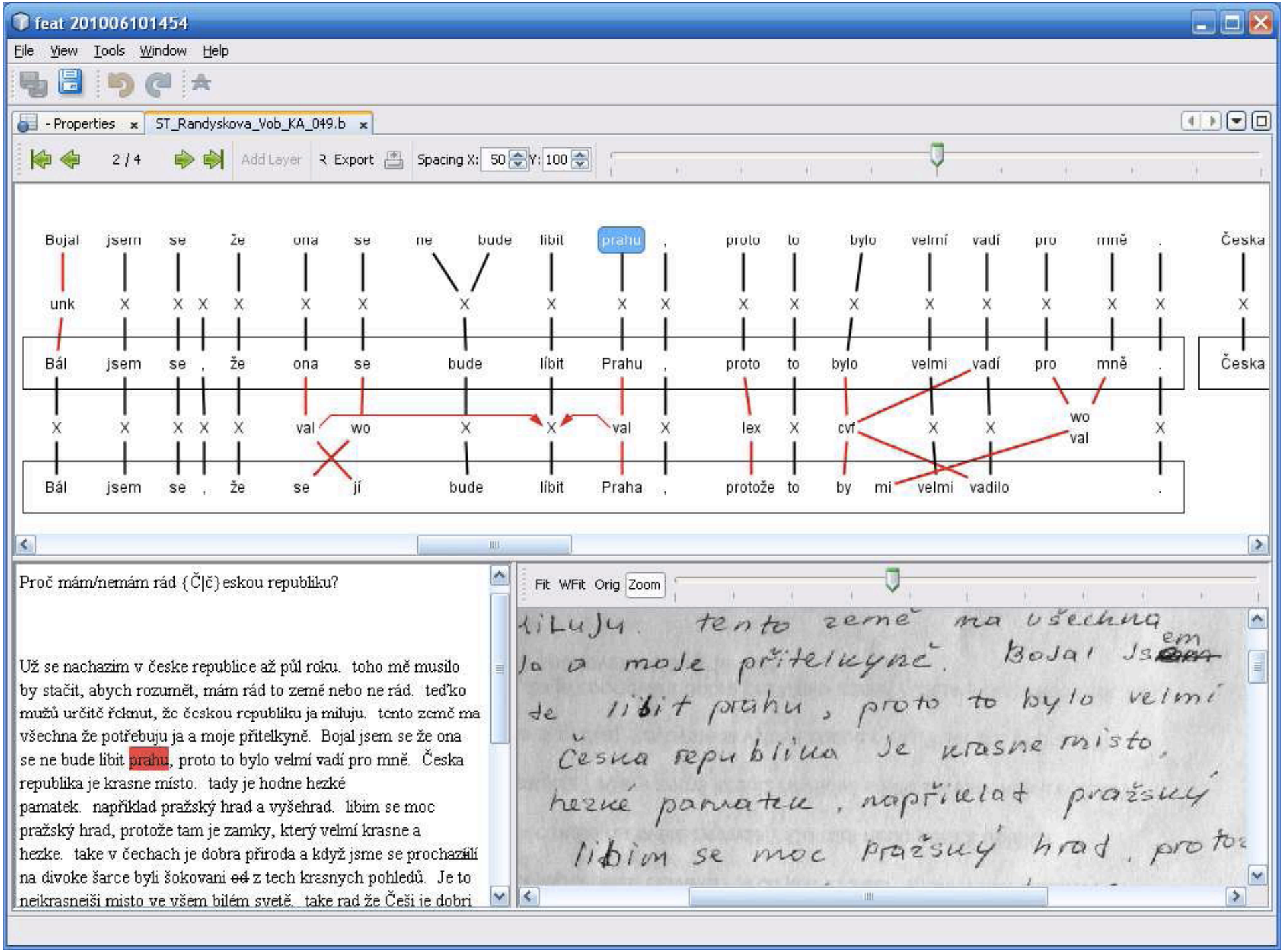

Bojal
jsme
*feared
AUX
incorInfl
Bál
jsme
agr
rflx
Bál
jsem
se
,
I was afraid




*that she would not like the famous city of Prague,*




*because I would be very unhappy about it.*


feat 201006101454
File View Tools Window Help
- Properties
ST_Randyskova_Vob_KA_049.b
2 / 4
Add Layer
Export
Spacing X: 50 Y: 100
Bojal jsem se že ona se ne bude libit prahu , proto to bylo velmi vadí pro mně . Česka
unk X X X X X X X X X X X X X X X X X X X
Bál jsem se , že ona se bude líbit Prahu , proto to bylo velmi vadí pro mně . Česka
X X X X X val wo X X val X lex X cvf X X wo val X
Bál jsem se , že se jí bude líbit Praha , protože to by mi velmi vadilo .
Proč mám/nemám rád (Č|č)eskou republiku?
Už se nachazim v ceske republice až půl roku. toho mě musilo by stačit, abych rozumět, mám rád to země nebo ne rád. teďko mužů určitč řcknut, žc čcskou rcpubliku ja miluju. tcnto zcmč ma všechna že potřebuju ja a moje přitelkyně. Bojal jsem se že ona se ne bude libit prahu, proto to bylo velmí vadí pro mně. Česka republika je krasne místo. tady je hodne hezké
památek. například pražský hrad a vyšehrad. libim se moc pražský hrad, protože tam je zamky, který velmí krasne a hezke. take v čechach je dobra příroda a když jsme se prochazili na divoke šarce byli šokovani ~~od~~ z tech krasnych pohledů. Je to neikrasneiši misto ve všem bilém svetě. take rad že Češi je dobri
Fit WFit Orig Zoom

# Chybová taxonomie

- 22 chybových značek doplňovaných manuálně
- 7 chybových značek doplňovaných automaticky

| | manuálně | automaticky | celkem |
|---|---|---|---|
| rovina 1 | 8 | 1 | 9 |
| rovina 2 | 11 | 6 | 17 |
| rovina 1 i 2 | 3 | 0 | 3 |
| celkem | 22 | 7 | 29 |

- další typy chyb určovaných automaticky
  - na základě porovnání původního tvaru s opraveným

## Anotační strategie

- anotovat gramatické a lexikální charakteristiky jazyka studentů
  - vzhledem ke spisovné normě

## Důvody

Anotace podle formálních kritérií se hodí jako základ pro:

- srovnání s jazykem rodilých mluvčích
- automatickou anotaci
- anotaci komunikativní kompetence, stylu apod.

# Obsah

# Evaluace

- taxonomie ověřována na dvojmo anotovaném vzorku 10 000 slov
- anotátorská shoda vyšší u chybových kategorií formálně přesně vymezených

| chybová značka | $\kappa$ |
|---|---|
| ***incorBase*** | 0,75 |
| ***agr*** | 0,54 |

Viz Štindlová (2011).

# Shoda mezi anotátory

9848 slov

| **značka** | **jen A1** | **jen A2** | **A1 i A2** | $\kappa$ |
|---|---|---|---|---|
| incor | 168 | 130 | 894 | 0,84 |
| incorStem | 167 | 165 | 559 | 0,75 |
| incorInfl | 173 | 130 | 250 | 0,61 |
| wbd | 14 | 21 | 45 | 0,72 |
| fw | 25 | 17 | 18 | 0,46 |
| agr | 82 | 99 | 110 | 0,54 |
| dep | 99 | 118 | 87 | 0,43 |
| neg | 11 | 9 | 9 | 0,47 |
| styl | 19 | 14 | 10 | 0,38 |
| lex | 107 | 131 | 74 | 0,37 |
| use | 60 | 74 | 19 | 0,21 |
| sec | 45 | 18 | 4 | 0,11 |

# Obsah

# Automatická anotace chybových textů

## Možnosti

- doplnění manuální anotace, morfosyntaktické značkování a POS
- předzpracování chybového textu
- plně automatická anotace

## Experimenty

- emendace pomocí automatického korektoru
  (zatím asi 82 % shody s anotátory)
- morfosyntaktické značkování chybového textu různými metodami, porovnání výsledků může vést k hypotéze o typu chyby
- automatická syntaktická analýza

# Obsah

# Výhled

## Podrobnější specifikace chybového aparátu

- na základě možností automatické detekce chyb
- na základě možností automatické emendace (spell-checker)
- na základě zkušeností s analýzami opřenými o CzeSL

# Obsah

## Děkujeme

dalším členům projektového týmu, zejména Karlu Šebestovi, Mileně Hnátkové, Tomáši Jelínkovi, Vladimíru Petkevičovi a Haně Skoumalové.

# Děkujeme za pozornost!