Building an Error-tagged Learner Corpus of Czech

Jirka Hana, Alexandr Rosen & Barbora Štindlová

Charles University, Prague & Technical University, Liberec

Institute of Formal and Applied Linguistics
Seminar of Formal Linguistics
April 22, 2013

Outline of the talk

1. Learner Corpora
2. CzeSL
3. Error Annotation of CzeSL
4. Evaluation
5. Postprocessing
6. Thanks

Outline of the talk

1 Learner Corpora
2 CzeSL
3 Error Annotation of CzeSL
4 Evaluation
5 Postprocessing
6 Thanks

Learner Corpora

- Include texts produced by learners of a foreign language
- Early 1990s: as a source of data for learners' dictionaries (e.g., *Longman Learner Corpus*)
- Used by authors of textbooks, methodologists and researchers in
  - Teaching *X* as a Foreign Language
  - Second Language Acquisition
- Deviant forms can be corrected and their error type identified
- There can be simultaneous deviations on multiple levels

Annotation of Learner Corpora

Learner corpora can be annotated in two independent ways:

Linguistic annotation

- Lemmatization, morphological tagging, syntactic structure, etc.
- On the original text or on the corrected text
- Usually automatic or semiautomatic

Error annotation

- Correcting and/or categorizing errors
- Diverse annotation systems
- Usually manual

Capturing errors

1. **Implicit** – errors are identified and corrected

- Pros:
  - faster training of annotators
  - faster process of annotation
- Cons:
  - results hard to search and analyze

2. **Explicit** – errors are identified and categorized

- Error categories (tags) reflect a specific theory

Outline of the talk

1 Learner Corpora

2 CzeSL

3 Error Annotation of CzeSL

4 Evaluation

5 Postprocessing

6 Thanks

CzeSL – A learner corpus of Czech

- The first large learner corpus of a Slavic language (*PiKUST* for Slovenian includes only 35KW)
- The first large learner corpus of Czech
- Part of the *AKCES* project – acquisition corpora of Czech (includes native speakers' spoken and written classroom language)

The parameters

- 2 MW (only a part annotated)
- 3 subcorpora by L1
  - Slavic language: Russian, Ukrainian, Polish
  - Other Indo-European language: German, English, French
  - Other language: Vietnamese, Arabic
- Written and spoken part
- All levels of proficiency according to CEFR
- Original documents are electronic and hand-written texts
- Elicited on various occasions in the class
- Metadata on learner and task (19 items)

Size

Size of various subcorpora (in thousand words, approximate)

	transcribed		annotated	doubly annotated
CzeSL – Foreigners				
spoken	11		0	
written	1,315		200	75
university	732		0	
ROMi – Czech Roma				
spoken	540	(270)	0	
written	450		170	110
native Czech				
spoken	1,046		?	
written	150		?	

Outline of the talk

1 Learner Corpora

2 CzeSL

3 Error Annotation of CzeSL

4 Evaluation

5 Postprocessing

6 Thanks

Workflow

- Acquisition
- Transcription
- Proofreading
- Conversion to PML
- Error annotation
- Revision
- *Adjudication*
- Postprocessing

Strategy

- Minimal correction
- Capture only grammatical and lexical characteristics of non-native language
- Relative to Literary Czech

Error Annotation of a Flective Language

Problems

- Inflection (nouns: 15 basic paradigms, subparadigms, subsub...)
- Derivation, agreement, word-order reflecting information structure, etc.

Solution

- Multilevel annotation scheme
- Combining manual and automatic annotation

Error Annotation of a Flective Language

Problems

- Inflection (nouns: 15 basic paradigms, subparadigms, subsub...)
- Derivation, agreement, word-order reflecting information structure, etc.

Solution

- Multilevel annotation scheme
- Combining manual and automatic annotation

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 českého profesora Smutneveselého a brzo dělá domácí úkol. Večeře Irena jde na prohlasku spolu z kamarádem, ale její bratr dělá nic. Jelo čeština je špatná, vím, že se vrátit ve Rusku budi studovat a pomalu myt podlahy.

Kamarad Ireny je Američan a chytry muž. On miluje Irenu a chce se vzít na ní, protože ona je hezká, taky chytra, rozumí mu a umí výborně

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.

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.

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.

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.

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.

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.

Multilevel Annotation Scheme

Level 0

- Original text (transcribed, self-corrections inlined)

Level 1

- Corrections disregarding word context
- Spelling, form of stems and endings
- Result: sequence of existing Czech forms

Level 2

- Remaining errors: syntactic, lexical, word-order, style, referential, negation, . . .
- Result: grammatically correct sentence

Bojal jsme se že ona se ne bude libila slavnou prahu, proto to bylo velmí vadí pro mně.

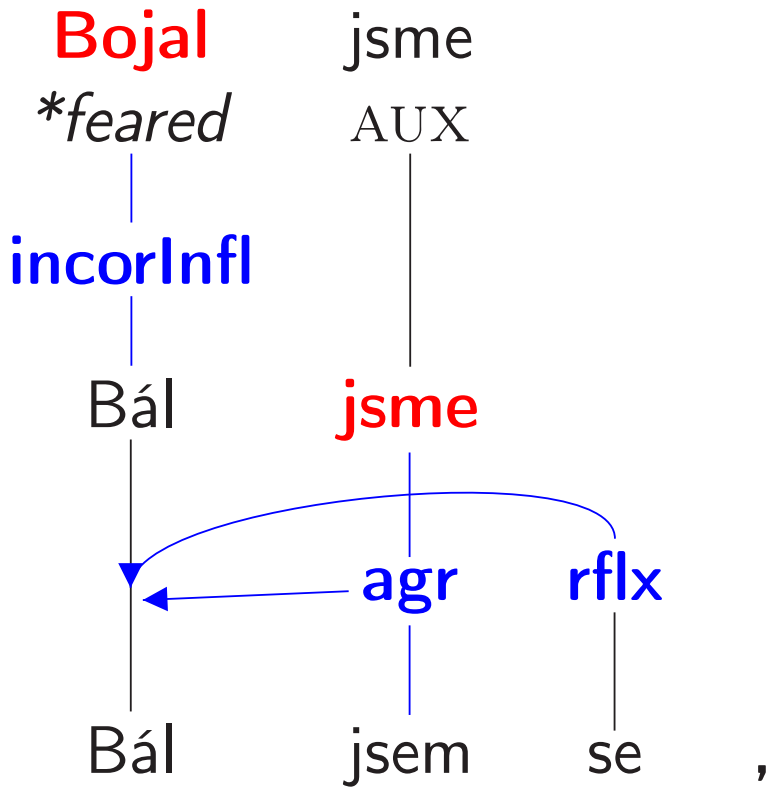
Bál jsem se, že se jí nebude líbit slavná Praha, protože to by mi velmi vadilo.

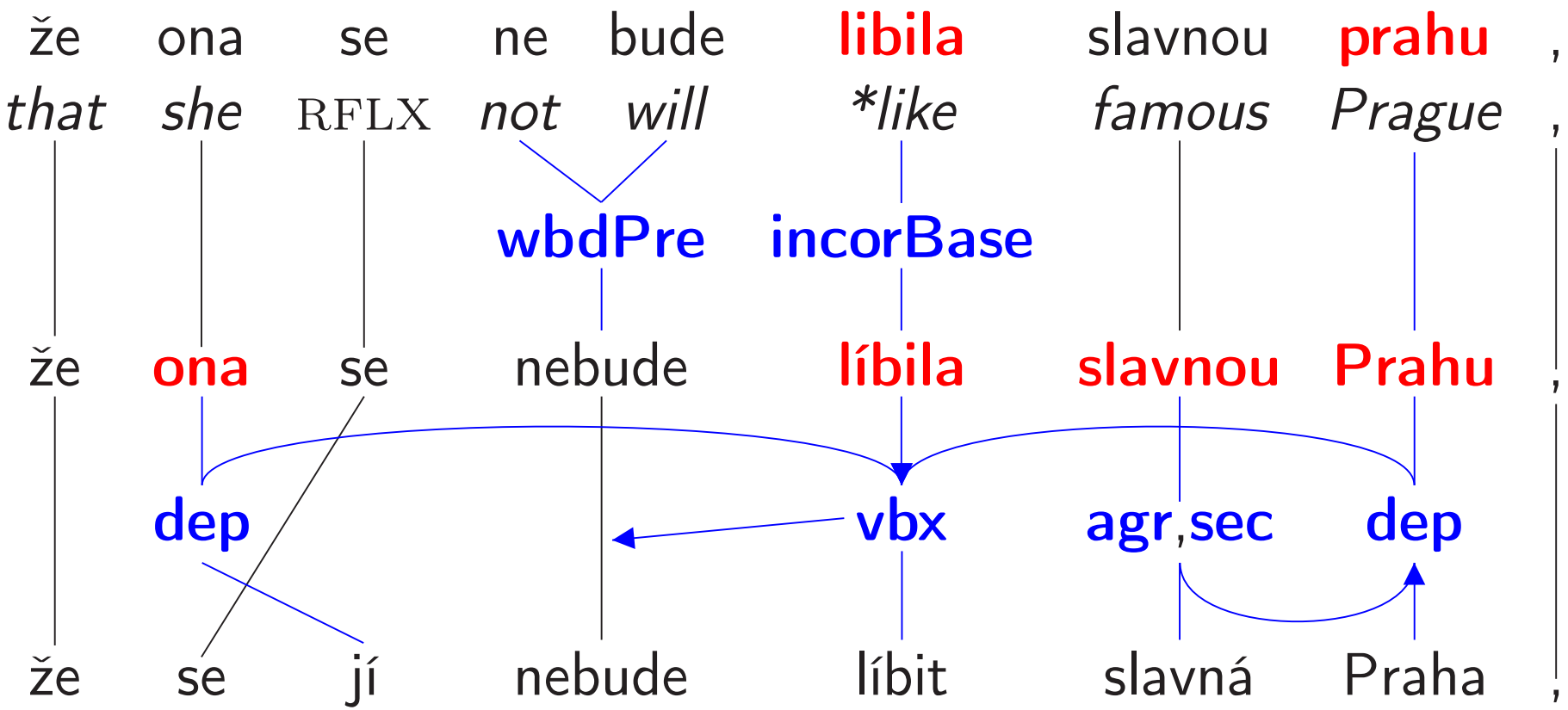
'I was affraid that she would not like the famous city of Prague, because I would be very unhappy about it.'

**Bojal jsme se že ona se ne bude libila slavnou prahu,
proto to bylo velmí vadí pro mně.**

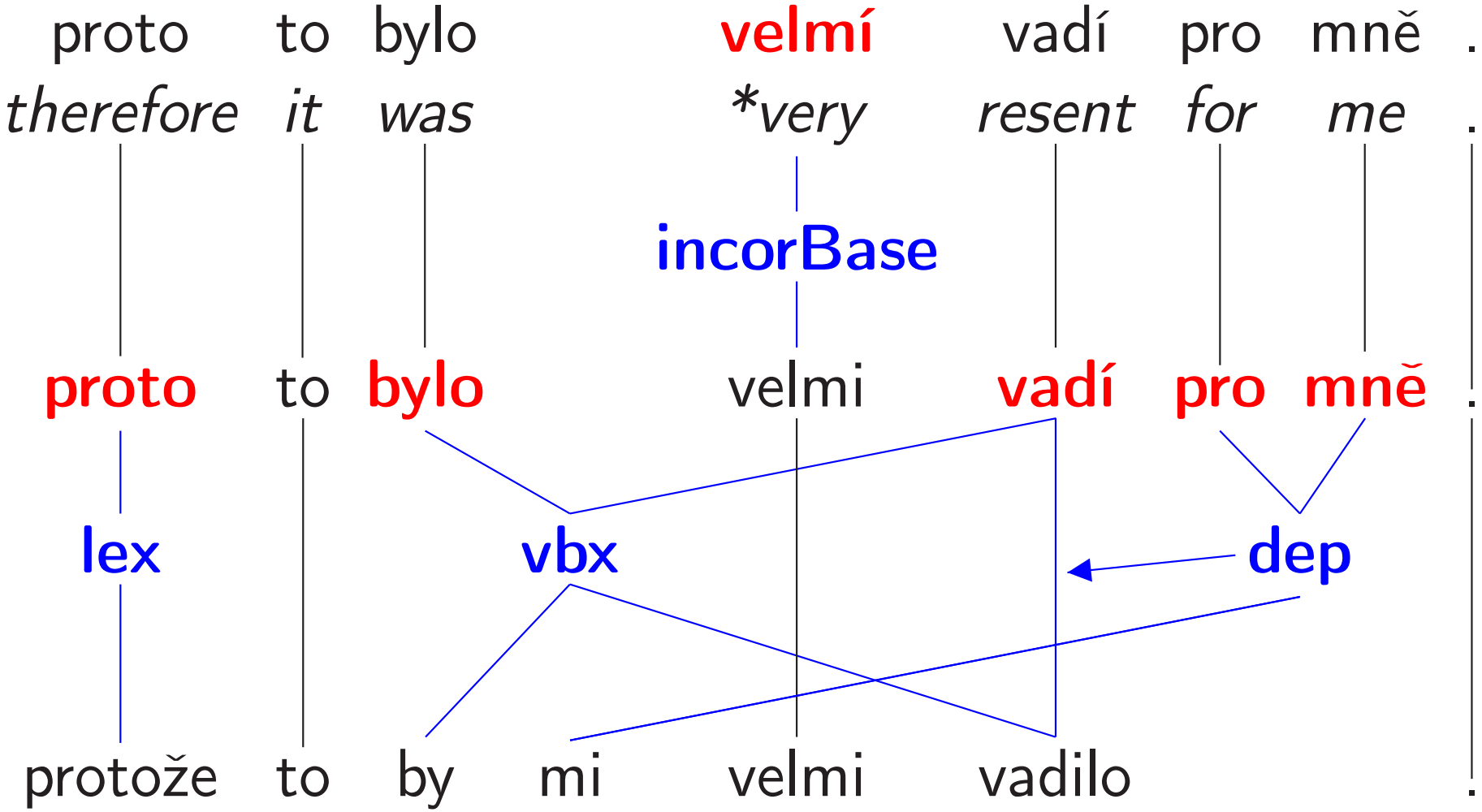
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

‘I was affraid that she would not like the famous city of Prague,
because I would be very unhappy about it.’

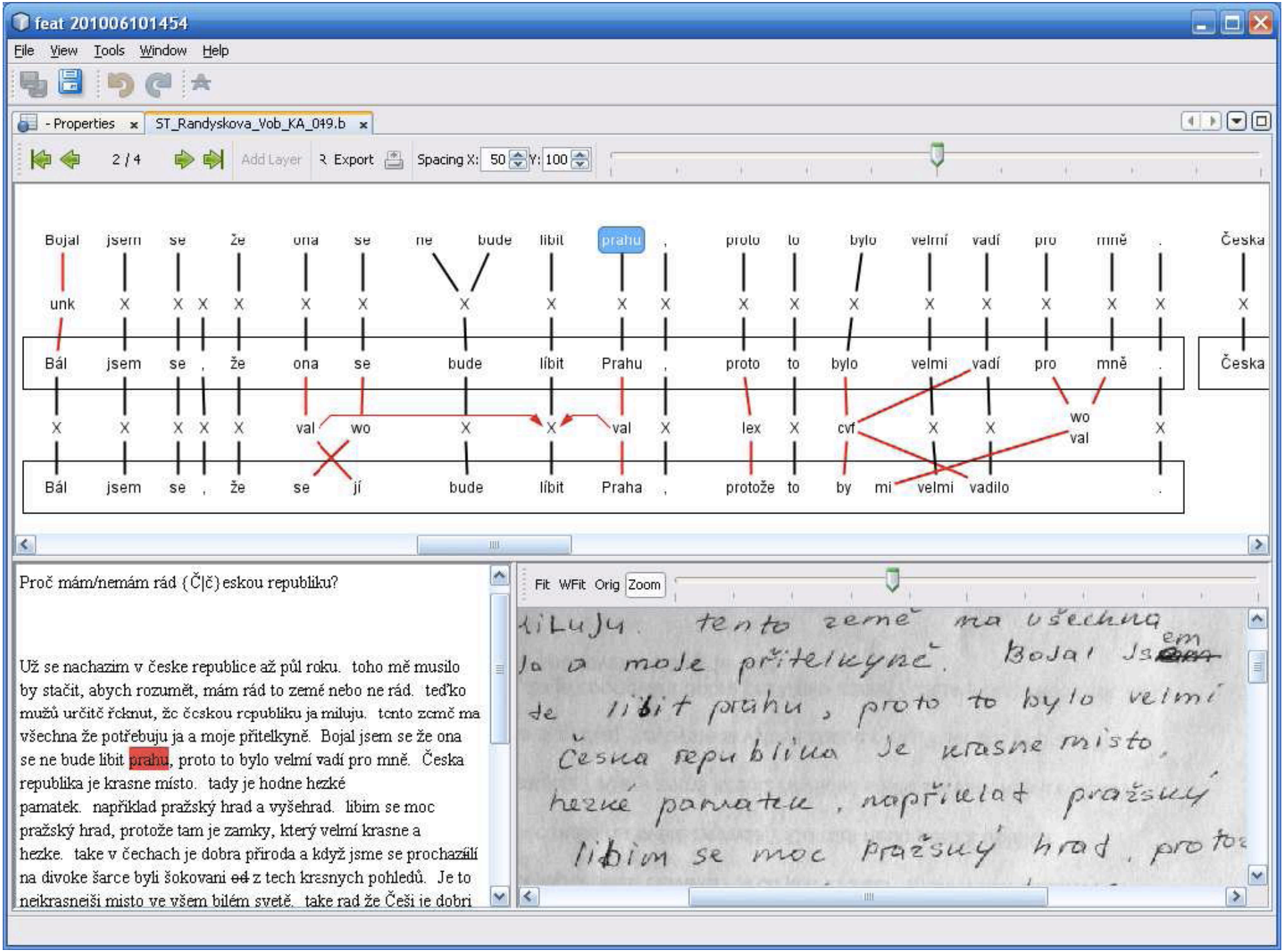

Bojal
jsme
*feared
AUX
incorInfl
Bál
jsme
agr
rflx
Bál
jsem
se
,
I was afraid



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

Error tags

- 22 error tags added manually
- 7 error tags added automatically

	manual	automatic	total
level 1	8	1	9
level 2	11	6	17
level 1 & 2	3	0	3
total	22	7	29

- Additional formal tags are added automatically by comparing original and corrected forms

incor			incorrect form	
	incorInfl	M	inflection error	L1
	incorBase	M	stem error	L1
	incorOther	A	other	L1
fw			foreign word, neologism, unidentifiable	
	fwFab	M	newly created “Czech” word	L1
	fwNc	M	foreign word	L1
	flex	M	inflection of fw	L1
wbd			word-boundary error	
	wbdPre	M	separate prefix, attached preposition	L1
	wbdComp	M	incorrectly separated/joined composites	L1
	wbdOther	M	other word-boundary errors	L1
styl			colloquial, bookish, regional expression	
	stylColl	M	colloquial expression	L1,L2
	stylOther	M	bookish, regional, slang expression	L1,L2
	stylMark	M	filler	L2
problem		M	problem	L1,L2

agr		M	agreement error	L2
dep		M	structural error	L2
ref		M	pronominal reference error	L2
vbx		M	complex verb error	L2
	cvf	A	analytical verb form error	L2
	mod	A	modal verb error	L2
	vnp	A	copula	L2
rflx		M	reflexive form error	L2
neg		M	negation error	L2
odd		A	extra word	L2
miss		A	missing word	L2
wo		A	word-order errror	L2
lex		M	lexical and idiomatic error	L2
use		M	incorrect use of a category	L2
sec		M	secondary error	L2
disr		M	word salad	L2

Outline of the talk

1 Learner Corpora

2 CzeSL

3 Error Annotation of CzeSL

4 Evaluation

5 Postprocessing

6 Thanks

Evaluation – Sample

- 67 texts, about 150 words each
- 9373 tokens
- CEFRL level A2–B1
- Various L1s
- 14 annotators, each text by two

A measure of IAA: Kappa

- $\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$

  $\Pr(a)$ – observed agreement
  $\Pr(e)$ – chance agreement
- κ = 1 – perfect agreement
- κ = 0 – random agreement
- κ > 0.4 – reasonable

Evaluation

Higher IAA on formally well defined tags

error tag	κ
incorBase	0.75
agr	0.54
styl	0.38

Inter-annotator agreement

9848 words

tag	only A1	only A2	A1 & A2	κ
incor	168	130	894	0.84
incorBase	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

Examples of high IAA

Agreement error

κ = 0.54

(1) Viděl malého Petra

(2) Viděl *malou Petra

Why not higher?

Different corrections

L0: *Věci budou *težki*
A1 – L1: *těžký*, L2: *těžké* + AGR
A2 – L1: *těžké*, L2: *těžké*

Examples of high IAA

Agreement error

κ = 0.54

(1) Viděl malého Petra

(2) Viděl *malou Petra

Why not higher?

Different corrections

L0: *Věci budou *težki*
A1 – L1: *těžký*, L2: *těžké* + AGR
A2 – L1: *těžké*, L2: *těžké*

Examples of low IAA

Lexical error

$\kappa = 0.37$

Due to semantic proximity of lexemes annotators disagree about the need for correction:

(3) *když se dívám na *?druhý/jiný kultury*
'when I look at other cultures'

On the other hand, some lexemes are distant enough and annotators agree about the need for for correction:

(4) **housenky/housky kupuju v pekařství*
'I buy caterpillars in the baker's shop'

Some reasons for low IAA

- Errors of type **lex** involve a high degree of subjective judgement, thus cannot aim at high IAA.
- Errors of type **sec** – highly formal specific, due to primary errors.

Evaluation – Conclusion

- Morphosyntactic errors are easy to formalize and lead to a high κ – incor, agr, dep
- Semantic errors depend on subjective judgement, should standard measures be applied?

Outline of the talk

1 Learner Corpora

2 CzeSL

3 Error Annotation of CzeSL

4 Evaluation

5 Postprocessing

6 Thanks

Postprocessing

- Formal error tags
- Morphology

Formal error tags

Error type	Error description	Example
`Cap0`	capitalization: incor. lower case	*evropě*/*Evropě*; *štědrý*/*Štědrý*
`Cap1`	capitalization: incor. upper case	*Staré*/*staré*; *Rodině*/*rodině*
`Vcd0`	voicing assimilation: incor. voiced	*stratíme*/*ztratíme*; *nabítku*/*nabídku*
`Vcd1`	voicing assimilation: incor. vcless	*zbalit*/*sbalit*; *nigdo*/*nikdo*
`VcdFin0`	word-final voicing: incor. voiceless	*kdyš*/*když*; *vztach*/*vztah*
`VcdFin1`	word-final voicing: incor. voiced	*přez*/*přes*; *pag*/*pak*
`Vcd`	voicing: other errors	*protoše*/*protože*; *hodili*/*chodili*
`Palat0`	missing palatalization (*k,g,h,ch*)	*amerikě*/*Americe*; *matkě*/*matce*
`Je0`	*je/ě*: incorrect *ě*	*ubjehlo*/*uběhlo*; *Nejvjetší*/*Největší*
`Je1`	*je/ě*: incorrect *je*	*vjeděl*/*věděl*; *vjeci*/*věci*
`Mne0`	*mě/mně*: incorrect *mě*	*zapoměla*/*zapomněla*
`Mne1`	*mě/mně*: incor. *mně, mňe, mňě*	*mněla*/*měla*; *rozumněli*/*rozuměli*
`ProtJ0`	protethic *j*: missing *j*	*sem*/*jsem*; *menoval*/*jmenoval*
`ProtJ1`	protethic *j*: extra *j*	*jse*/*se*; *jmé*/*mé*
`ProtV1`	protethic *v*: extra *v*	*vosm*/*osm*; *vopravdu*/*opravdu*
`EpentE0`	*e* epenthesis: missing *e*	*domček*/*domeček*
`EpentE1`	*e* epenthesis: extra *e*	*rozeběhl*/*rozběhl*; *účety*/*účty*

Morphology

- L2 (correct Czech): disambiguated lemma + tag
- L1: morphological analysis

- If possible, disambiguated lemma+tag is transfered from L2 to L1

- If forms are different, but L2 lemma is among L1 lemmas:
  use L2 lemma for L1 L1=*má* 'has' or 'my'

  L2=*mou* 'my': use the lemma *můj* 'my' on L1

Outline of the talk

1 Learner Corpora

2 CzeSL

3 Error Annotation of CzeSL

4 Evaluation

5 Postprocessing

6 Thanks

Thanks to...

... other members of the team, esp.:

Barbora Štindlová, Tomáš Jelínek, Svatava Škodová, Karel Šebesta, Vladimír Petkevič, Hana Skoumalová, Milena Hnátková, Jan Štěpánek, Zuzanna Bedřichová, Kateřina Šormová

... the sponsors:

- The European Social Fund and the Czech government: *Education for Competitiveness – Innovation in Education in the Field of Czech as a Second Language* (CZ.1.07/2.2.00/07.0259)
- Large Research, Development and Innovation Infrastructures: *The Czech National Corpus* (LM2011023)

... and you!