Building and using corpora of non-native Czech

Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Faculty of Arts
Charles University in Prague

SloNLP 2016
Tatranské Matliare
17–19 September 2016

Outline of the talk

1. About learner corpora
2. CzeSL – a corpus of Czech as a Second Language
3. CzeSL without metadata: CzeSL-plain and CzeSL-man v. 0
4. Automatic error annotation: CzeSL-SGT
5. Manual annotation and metadata: CzeSL-man v. 1
6. Issues and lessons learnt

Outline of the talk

Learner Corpora

- Include texts produced by learners of a foreign language (L2)
- Early 1990s: as a source of data for learners' dictionaries (e.g., *Longman Learner Corpus*)
- 2002: *International Corpus of Learner English (ICLE)* – University of Louvain
- Used by authors of textbooks, methodologists and researchers in L2 acquisition
- Deviant forms can be corrected and their error type identified
- Many specifics in comparison to standard corpora

Using a learner corpus

- To describe levels of progress in learners' interlanguage
- To identify an optimal order and method of teaching grammar
- To research L1 influence
- To identify overuse and underuse of linguistic items in learner language
- To identify features responsible for the 'foreign sound'

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, in-line
- Sometimes multi-layered, with more error types and/or corrections at different linguistic levels, including intelligibility score

Corpus	Size (MW)	L1	L2	Level	Medium	Annotation
ICLE	3.7	26	en	advanced	written	part
CLC	45	130	en	all	written	part
LINDSEI	0.8	11	en	advanced	spoken	part
PELCRA	0.5	pl	en	all	written	part
USE	1.2	sv	en	advanced	written	no
HKUST	25	zh	en	advanced	written	part
CHUNGDAHM	131	ko	en	all	written	part
JEFLL	0.7	jp	en	beginners	written	part
MELD	1	16	en	advanced	written	no
MICASE	1.8	var	en	advanced	spoken	no
NICT JLE	2	jp	en	all	spoken	part
RusLTC	1.5	ru	en	advanced	written	no
FALKO	0.3	5	de	advanced	written	part
FRIDA	0.3	var	fr	med-adv	written	part
FLLOC	2	en	fr	all	spoken	no
PiKUST	0.04	18	sl	advanced	written	yes
ASU	0.5	var	no	advanced	written	no
TUFS	0.6	var	jp	all	written	no
MERLIN	0.1	var	de,cs,it	all	written	yes

Outline of the talk

CzeSL – **Cze**ch as a **S**econd **L**anguage

- A part of *AKCES* – an umbrella project, various funding
- Groups:
  - Native learners of Czech
  - Romani ethnolect of Czech
  - **Non-native learners of Czech**
- **Written**/spoken language

Non-native learners

- Transcribed hand-written essays
- `http://utkl.ff.cuni.cz/learncorp/`
- *CzeSL-plain* – also Romani ethnolect and native (2.3M tokens)
- *CzeSL-SGT* – automatic error annotation, tagged (1.1M tokens)
- *CzeSL-MAN* – manual annotation, tagged, parsed (124K tokens)
- L1 groups:
  - Slavic: Russian, Ukrainian, Polish, ...
  - Other Indo-European: German, English, French, ...
  - Non-Indo-European: Vietnamese, Chinese, Arabic, ...
- All levels of proficiency according to CEFR
- Metadata on the learner and the task (30 items)

Workflow

- Acquisition
- Transcription
- Proofreading
- Conversion to PML
- *Manual error annotation*
- *Revision*
- *Adjudication*
- Automatic linguistic and/or error annotation

Available releases of CzeSL

	Non-native		Ethno	TOTAL	Annot	Meta
	Essays	Theses				
CzeSL-plain	1315	732	428	2475	no	no
CzeSL-SGT	1147			1147	auto	yes
CzeSL-man v.0, a1	134		192	326	manual	no
CzeSL-man v.0, a2	59		149	208	manual	no
CzeSL-man v.1	134			134	manual	yes

(Sizes in thousands of tokens)

Outline of the talk

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.**

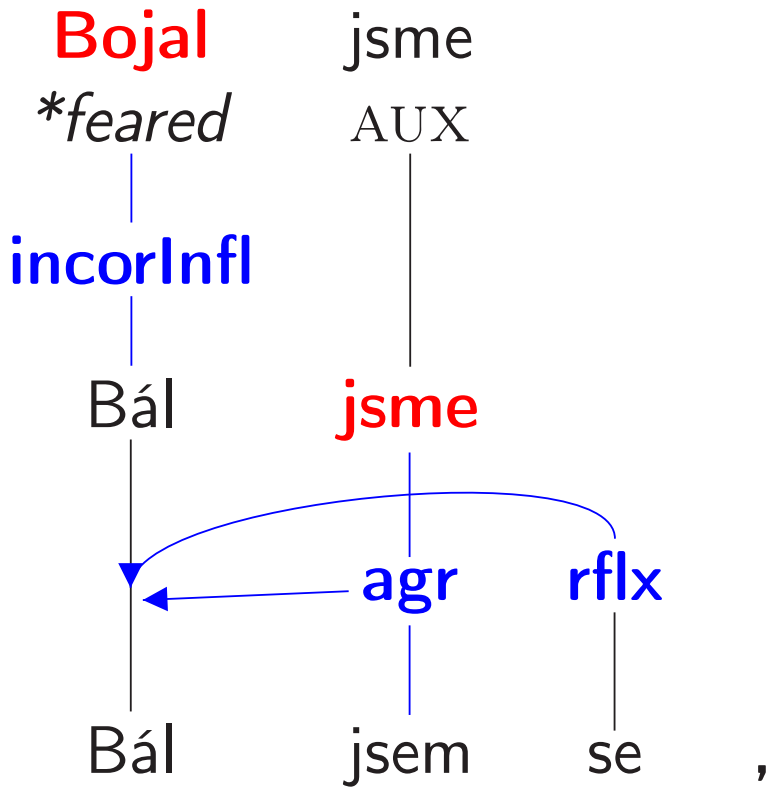
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

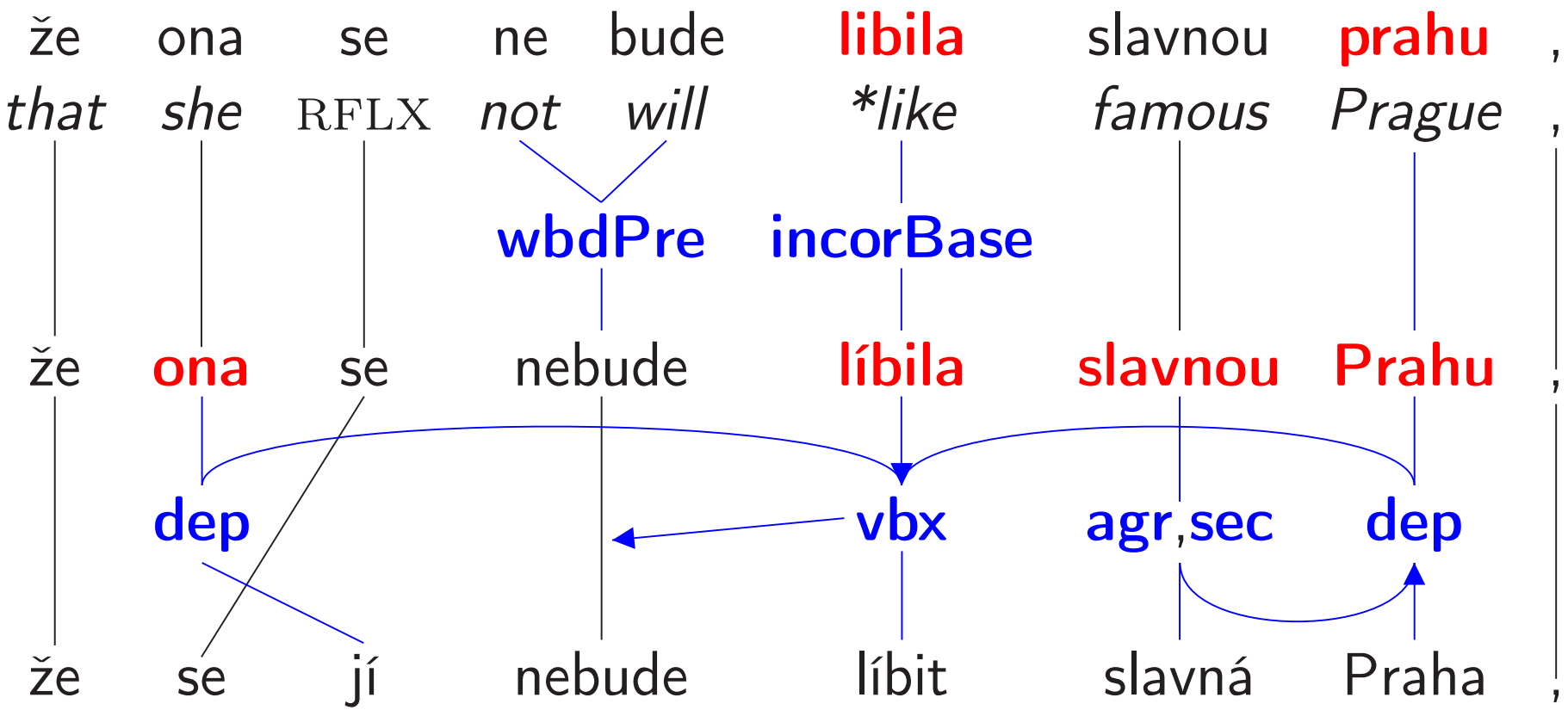
‘I was affraid that she would not like the famous city of Prague,
because I would be very unhappy about it.’

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

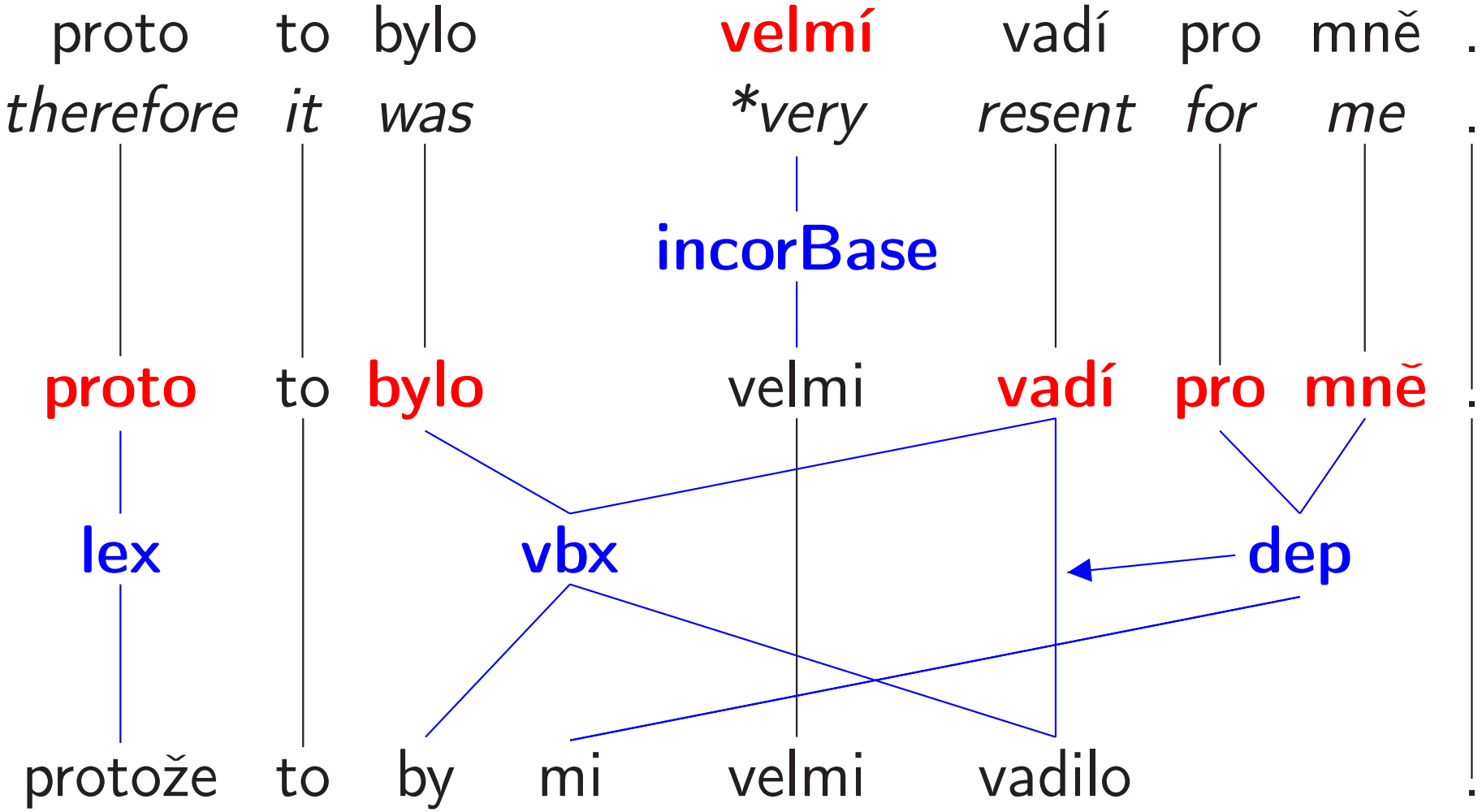
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

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

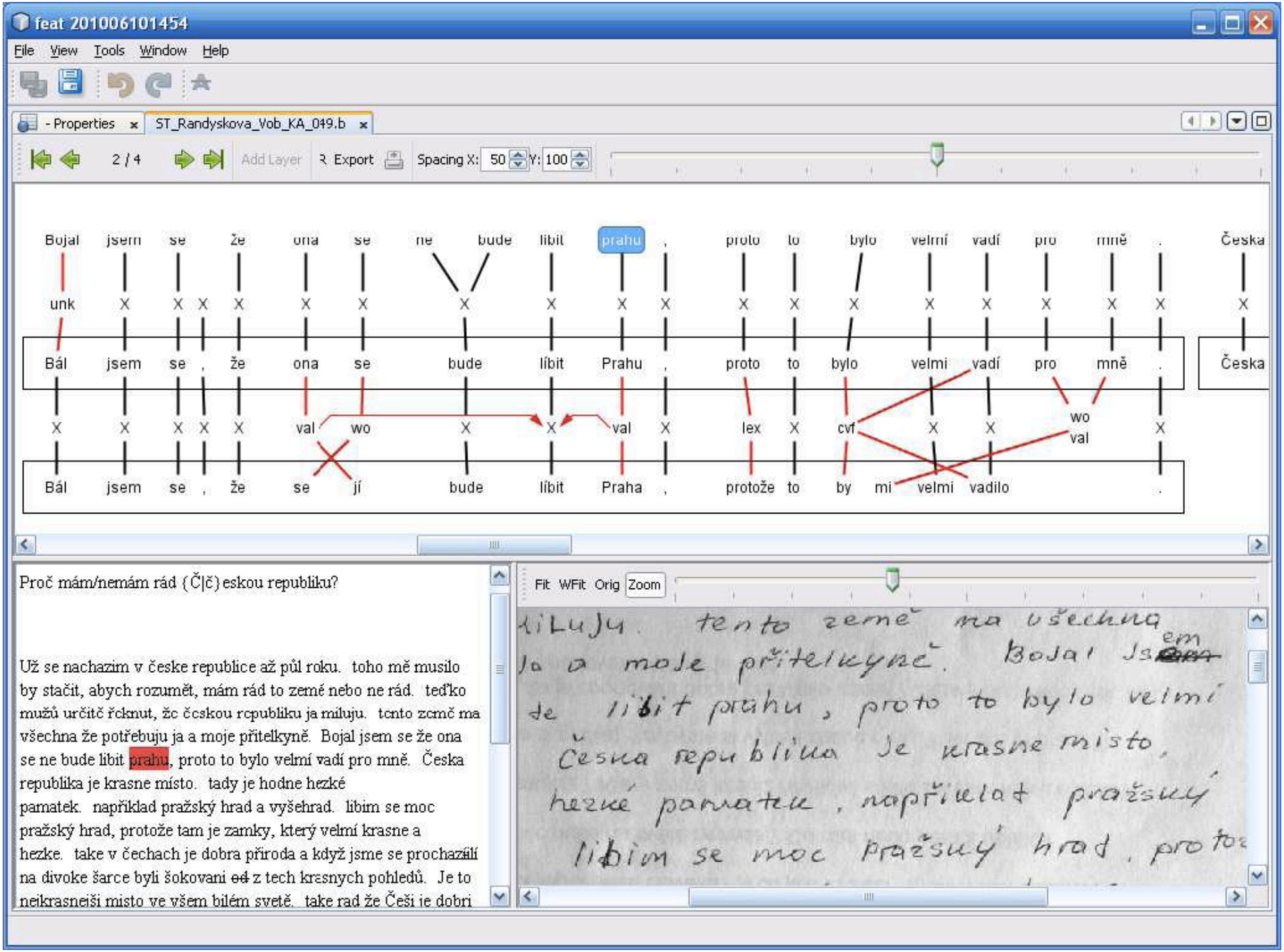

Bojal
jsme
*feared
AUX
incorInfl
Bál
jsme
agr
rflx
Bál
jsem
se
,
I was afraid



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

Outline of the talk

CzeSL-SGT

- **Cze**ch as a **S**econd **L**anguage with **S**pelling, **G**rammar and **T**ags
- With metadata about the text and the author
- With automatic linguistic and error annotation
  - correction
  - tagging and lemmatization
  - error labels
- Searchable from the interface of the Czech National Corpus: `http://kontext.korpus.cz`
- Downloadable from the LINDAT data repository (*AKCES 5*): `http://www.lindat.cz`[1]

[1] `http://hdl.handle.net/11234/1-162`

Annotation

- If possible, each word form is tagged by a standard tagger[2] with:
  - word class
  - morphological categories
  - base form (lemmas)
- Forms detected as incorrect are corrected by a stochastic spelling and grammar checker, targeting even some 'real word' errors[3]
- The corrected text is re-tagged
- Original and corrected forms are compared and error labels, based on applicable formal criteria, are assigned[4]
- All the annotation is assigned automatically

[2] [Votrubec(2006)]
[3] [Richter(2010), Richter et al.(2012)]
[4] [Jelínek et al.(2012)]

The tools: Spell/Grammar-checker

Korektor[5]

- Combines rule-based morphology with a stochastic model
- Trained on native texts (Prague Dependency Treebank)[6]
- Ranked suggestions with a correction type: spelling or grammar
- Single words only

[5] [Richter et al.(2012)],[Ramasamy et al.(2015)]
[6] [PZK(2005)]

Evaluation of the automatic correction

Korektor

- The sample: 67 texts, 9373 tokens, 7995 words
- Evaluated on a manually and doubly annotated subset of CzeSL
- Only where both annotators agree
- Results for ill-formed tokens: 82%[a]

[a] [Štindlová et al.(2012)]

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*

A sentence with “spelling” errors

(1) Tén pes míluje svécho kamarada – člověka.
Ten pes miluje svého kamaráda – člověka.
‘That dog loves his friend – the man.’

word	lemma	tag	word1	lemma1	tag1	gs	err
Tén	Tén	X@	Ten	ten	PDYS1	S	Quant1
pes	pes	NNMS1	pes	pes	NNMS1		
míluje	míluje	X@	miluje	milovat	VB-S-3P	S	Quant1
svécho	svécho	X@	svého	svůj	P8MS4	S	Voiced
kamarada	kamarada	X@	kamaráda	kamarád	NNMS4	S	Quant0
-	-	Z:	-	-	Z:		
člověka	člověk	NNMS2	člověka	člověk	NNMS4		
.	.	Z:-	.	.	Z:		

A sentence with “real-word” errors

(2) Nejakij muž spí v postele.
Nějakej muž spí v posteli.
‘Some guy is sleeping in the bed.’

word	**lemma**	**tag**	**word1**	**lemma1**	**tag1**	**gs**	**err**
Nejakij	Nejakij	X@	Nějakej	nějaký	PZYS1-6	S	Caron0
muž	muž	NNMS1	muž	muž	NNMS1		
spí	spát	VB-S---3P	spí	spát	VB-S---3P		
v	v	RR--4	v	v	RR--6		
postele	postel	NNFP4	posteli	postel	NNFS6	G	SingCh
.	.	Z:	.	.	Z:		

Metadata

Most texts are equipped with metadata about the author and the text.

15 items about the author:

- sex
- age
- L1
- CEFR level of proficiency in Czech
- duration and method of study
- length of stay in Czechia
- knowledge of Czech among family members
- ...

Metadata, cont'd

15 items about the text:

- date
- time limit
- word count
- topic
- genre
- dictionary/textbook allowed
- exam?
- ...

Number of texts by language group and proficiency level in *CzeSL-SGT*

	S	IE	nIE	unknown	Σ
A1	1783	199	622	5	2609
A1+	283	21	11	0	315
A2	1348	269	480	1	2098
A2+	403	54	113	0	570
B1	929	195	357	0	1481
B2	523	115	107	0	745
C1	82	17	24	0	123
C2	0	1	0	0	1
unknown	291	27	33	324	675
Σ	5642	898	1747	330	8617

Searching the corpus

Global conditions in a CQL query

- 1:[] & 1.lemma != 1.lemma1
- 1:[] & 1.c != 1.c1

Životní styl , kultura , služby v ČR a v	me /mé/X/6	zemi Rozdíl životního stylu mězy Čechami
styl , kultura , služby v ČR a v me	zemi /zemi/3/6	Rozdíl životního stylu mězy Čechami a Rus
tura , služby v ČR a v me zemi Rozdíl	životního /životního/2/2	stylu mězy Čechami a Ruskem je moc velk
by v ČR a v me zemi Rozdíl životního	stylu /stylu/2/2	mězy Čechami a Ruskem je moc velky , tak
ČR a v me zemi Rozdíl životního stylu	mězy /mezi/-/7	Čechami a Ruskem je moc velky , také jak
me zemi Rozdíl životního stylu mězy	Čechami /Čechami/7/7	a Ruskem je moc velky , také jak a kultura
ozdíl životního stylu mězy Čechami a	Ruskem /Ruskem/7/7	je moc velky , také jak a kultura , a
tylu mězy Čechami a Ruskem je moc	velky /velký/-/1	, také jak a kultura , a služby . V
je moc velky , také jak a kultura , a	služby /služby/4/4	. V minulem roku , když ještě jsem bydlila v
, také jak a kultura , a služby . V	minulem /minulém/7/6	roku , když ještě jsem bydlila v Rusku , chtě
é jak a kultura , a služby . V minulem	roku /roku/2/6	, když ještě jsem bydlila v Rusku , chtěla po
lyž ještě jsem bydlila v Rusku , chtěla	pojet /počet/-/4	studovat v Prahu , protože myslila , že tady
noc příjemná počasí , tzn. že ne moc	hladno /Kladno/-/1	nebo horko , jak rozdílnost měho rodného
n. že ne moc hladno nebo horko , jak	rozdílnost /rozdílnost/4/1	měho rodného města , také jsem myslila , ž
oc hladno nebo horko , jak rozdílnost	měho /mého/-/2	rodného města , také jsem myslila , že ČR j
dno nebo horko , jak rozdílnost měho	rodného /rodného/2/2	města , také jsem myslila , že ČR je bezpeč
horko , jak rozdílnost měho rodného	města /města/2/2	, také jsem myslila , že ČR je bezpečná ,
sem myslila , že ČR je bezpečná , lidé	v /v/4/4	ně dodržujou zákony a policia a všechny o
m myslila , že ČR je bezpečná , lidé v	ně /ně/4/4	dodržujou zákony a policia a všechny ochr

Outline of the talk

CzeSL-man v. 1 and *CzeSL-SGT* compared

	CzeSL-SGT	*CzeSL-man v. 1*
Texts	8,600	645
Sentences	111K	11K
Words	958K	104K
Tokens	1,148K	128K
Doubly annotated		46%
Different authors	1,965	262
Different L1s	54	32
Proficiency levels	A1–C2	A1–C1
Women/Men	5:3	3:2
Words per text	100–200	100–200

Number of texts by language group and proficiency level in *CzeSL-man v. 1*

	S	IE	nIE	unknown	Σ
A1	49	6	4		59
A1+			3		3
A2	18	26	67		111
A2+	81	9	59		149
B1	123	26	30		179
B2	102	11	15		128
C1	10		2		12
unknown				4	4
Σ	383	78	180	4	645

Outline of the talk

The choice of texts

Choice of texts

CzeSL-plain and *CzeSL-man v. 0* include some *ROMi* texts

Metadata

Metadata: *CzeSL-plain* and *CzeSL-man v. 0* lack them

Two-level annotation

Two-level annotation: *SeLaQ* cannot display it in a graphical format

- *Manatee*: in-line annotation using embedded XML elements

přečístse $\Longrightarrow$ *přečíst si*

<**err1** type=wbdJoin>
přečístse
</**err1**>
<corr1 type=wbdJoin>
přečíst
<**err2** type=lex>
se
</**err2**>
<corr2 type=lex>
si
</corr2>
</corr1>

Thanks to...

Jirka Hana
Tomáš Jelínek
Barbora Štindlová
Vojtěch Kovář
Pavel Procházka
Hana Skoumalová
...

... and you!

References I

Jelínek, T., Petkevič, V., Rosen, A., & Skoumalová, H. (2012).
Czech treebanking unlimited.
In J. Hajič, K. D. Smedt, M. Tadić, and A. Branco, editors, *Proceedings of the META-RESEARCH Workshop on Advanced Treebanking, LREC 2012*, pages 37–44, Istanbul, Turkey. ELRA, European Language Resources Association.

PZK (2005).
Pražský závislostní korpus.
Ústav formální a aplikované lingvistiky MFF UK, Praha.
Verze 2.0, http://ufal.mff.cuni.cz/pdt/.

References II


Ramasamy, L., Rosen, A., & Straňák, P. (2015).
Improvements to Korektor: A case study with native and non-native Czech.
In J. Yaghob, editor, *ITAT 2015: Information technologies – Applications and Theory / SloNLP 2015*, pages 73–80, Prague. Charles University in Prague.


Richter, M. (2010).
An Advanced Spell Checker of Czech.
Master's thesis, Faculty of Mathematics and Physics, Charles University, Prague.

References III

Richter, M., Straňák, P., & Rosen, A. (2012).
Korektor – a system for contextual spell-checking and diacritics completion.
In *Proceedings of COLING 2012: Posters*, pages 1019–1028, Mumbai, India. The COLING 2012 Organizing Committee.

Votrubec, J. (2006).
Morphological tagging based on averaged perceptron.
In *WDS'06 Proceedings of Contributed Papers*, pages 191–195, Praha, Czechia. Matfyzpress, Charles University.

Štindlová, B., Rosen, A., Hana, J., & Škodová, S. (2012).
CzeSL – an error tagged corpus of Czech as a second language.
In P. Pęzik, editor, *Corpus Data across Languages and Disciplines*, volume 28 of *Łódź Studies in Language*, pages 21–32, Frankfurt am Main. Peter Lang.