Akviziční korpusy

Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofické fakulty Univerzity Karlovy v Praze

Korpusový seminář
17. března 2016

Osnova

1. Akviziční korpusy a jejich anotace
2. Learner Corpora of Czech: Merlin and CzeSL
3. Error Annotation of CzeSL
4. An automatically annotated corpus – CzeSL-SGT
5. Searching the corpus
6. Theoretical issues

Osnova

Akviziční korpusy

- Pro studium osvojování a výuky mateřského i cizího (druhého) jazyka
- Doklady o užívání jazyka mluvčími, kteří si jazyk (dosud) neosvojili na úrovni dospělého rodilého mluvčího
- Psané i mluvené
- Zaměřené na rodilé i nerodilé mluvčí
- Někdy s chybovou anotací, která vyznačuje odchylky od běžného úzu nebo normy
- Někdy longitudinální nebo kvazilongitudinální
- ČNK: SCHOLA2010, SKRIPT2012, CzeSL-plain, CzeSL-SGT[1]

[1] http://akces.ff.cuni.cz,http://utkl.ff.cuni.cz/learncorp/, https://www.facebook.com/novacds/videos/vb.501336856549038/978415935507792/?type=2&theater

Žákovské korpusy (learner corpora)

- Obsahují texty nerodilých mluvčích, kteří se daný jazyk učí.
- Většinou psané školní práce, často v rámci zkoušky, někdy i mluvené žákovské korpusy.
- Důraz na metadata: věk, mateřský jazyk, úroveň znalosti cílového jazyka apod.
- Od 1990– jako materiál pro slovníky určené studentům angličtiny (např. *Longman Learner Corpus*)
- 2002: *International Corpus of Learner English (ICLE)*
  – Univerzita v Nové Lovani
- Pro autory učebnic, metodology, jazykovědce
- Odchylky od úzu/standardu lze opravovat a určit typ chyby
- Odchylky mohou být na více rovinách současně

Korpus	M slov	L1	L2	Úroveň	Médium	Anotace
ICLE	3	26	en	pokr.	psaný	část.
CLC	35	130	en	vše	psaný	část.
LINDSEI	0.8	11	en	pokr.	mluv.	část.
PELCRA	0.5	pl	en	vše	psaný	část.
USE	1.2	sv	en	pokr.	psaný	ne
HKUST	25	zh	en	pokr.	psaný	část.
CHUNGDAHM	131	ko	en	vše	psaný	část.
JEFLL	0.7	jp	en	zač.	psaný	část.
MELD	1	16	en	pokr.	psaný	ne
MICASE	1.8	různ.	en	pokr.	mluv.	ne
NICT JLE	2	jp	en	vše	mluv.	část.
FALKO	0.3	5	de	pokr.	psaný	část.
FRIDA	0.2	různ.	fr	stř.-pokr.	mluv.	část.
FLLOC	2	en	fr	vše	mluv.	ne
PiKUST	0.04	18	sl	pokr.	psaný	ano
ASU	0.5	různ.	no	pokr.	psaný	ne
TUFS	0.6 M znaků	různ.	jp	vše	psaný	ne

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 distinguish universal errors from errors due to learner's L1
- 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

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

A sample (manual) error annotation – CzeSL[2]

incor			incorrect form	
	incorInfl	M	inflection error	T1
	incorBase	M	stem error	T1
	incorOther	A	other	T1
fw			foreign word, neologism, unidentifiable	
	fwFab	M	newly created “Czech” word	T1
	fwNc	M	foreign word	T1
	flex	M	inflection of fw	T1
wbd			word-boundary error	
	wbdPre	M	separate prefix, attached preposition	T1
	wbdComp	M	incorrectly separated/joined composites	T1
	wbdOther	M	other word-boundary errors	T1
styl			colloquial, bookish, regional expression	
	stylColl	M	colloquial expression	T1,T2
	stylOther	M	bookish, regional, slang expression	T1,T2
	stylMark	M	filler	T2
problem		M	problem	T1,T2

[2][Štindlová et al.(2013), Rosen et al.(2014a)]

agr		M	agreement error	T2
dep		M	structural error	T2
ref		M	pronominal reference error	T2
vbx		M	complex verb error	T2
	cvf	A	analytical verb form error	T2
	mod	A	modal verb error	T2
	vnp	A	copula	T2
rflx		M	reflexive form error	T2
neg		M	negation error	T2
odd		A	extra word	T2
miss		A	missing word	T2
wo		A	word-order errror	T2
lex		M	lexical and idiomatic error	T2
use		M	incorrect use of a category	T2
sec		M	secondary error	T2
disr		M	word salad	T2

Automatic annotation?

- For many native languages, reasonably reliable annotation tools are available.
- Non-native language is often annotated manually, but this is not realistic for larger volumes.
- Can methods and tools developed for native language help?

Automatic annotation of learner corpora

- NLP for learner language:[3]
  - tutoring systems
    (Intelligent Computer-Assisted Language Learning – ICALL)[4]
  - automated scoring in language testing
  - analysis and annotation of learner corpora
- Linguistic annotation: lemmatization, tagging, (shallow) parsing[5]
- Error annotation

[3] [Meurers(2013)]
[4] [Dickinson & Herring(2008)]
[5] [Nagata et al.(2011), Dickinson & Ragheb(2009), Krivanek & Meurers(2014)]

Osnova

Merlin

- Learner corpus of Czech, German, and Italian[6]
- To build a platform matching CEFR levels with language phenomena specific to the level
- Funded by the EU Lifelong Learning Programme, 2012–2014
- Czech: 64.5K words, CEFR levels A1–C1
- Tagged, parsed, on-line searchable

[6] [Boyd et al.(2014)], http://www.merlin-platform.eu

AKCES – Acquisition corpora of Czech[7]

- An umbrella project, various funding
- Faculty of Arts, Charles University in Prague
- Project head: Karel Šebesta,
  Institute of Czech Language and Theory of Communication
- Groups:
  - Native learners
  - Learners growing up in socially excluded communities, mostly with Romani background
  - Non-native learners
- Written/spoken language

[7] [Šebesta(2012)], http://akces.ff.cuni.cz, http://utkl.ff.cuni.cz/learncorp/

- Searchable:
  - CNC – http://kontext.korpus.cz
  - ITCL – http://chomsky.ruk.cuni.cz:5125
- Downloadable:
  - LINDAT – http://lindat.mff.cuni.cz
  - License – Creative Commons BY-(NC-)ND 3.0

- Native learners, elementary and secondary school
  - Speech: *SCHOLA 2010* ≈ *AKCES 2* (1M tokens)
  - Essays: *SKRIPT 2012* ≈ *AKCES 1* (0.7M tokens), *SKRIPT-SGT* – in prep., also Roma learners
- Roma learners
  - Speech: *ROMi 1.0* (1.5M words)
  - Essays: *AKCES 4* (300K words)
- Non-native learners (essays)
  - *CzeSL-plain* ≈ *AKCES 3* – also Roma and native (2.3M tokens)
  - *CzeSL-SGT* ≈ *AKCES 5* – automatic annotation (1.1M tokens)
  - *CzeSL-MAN* – manual annotation (288K/48K tokens, SeLaQ beta)

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

- Approx. 1 MW, transcribed hand-written essays
- 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)

Sizes and proportions

Texts	8.6K
Sentences	111K
Words	958K
Tokens	1,148K
Different authors	1,965
Different native languages	54
Proficiency levels	A1–C2
Age	9–76
Women/Men	5/3 KW
Words per text	100–200

Language groups and proficiency levels[8]

CEFR (Czech)	L1 group				
	Slavic	IndEur	non-IndEur	?	Total
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
?	291	27	33	324	675
Total	5642	898	1747	330	8617

[8] More statistics on http://utkl.ff.cuni.cz/~rosen/public/sgt_counts_by_meta_en.html

Osnova

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.

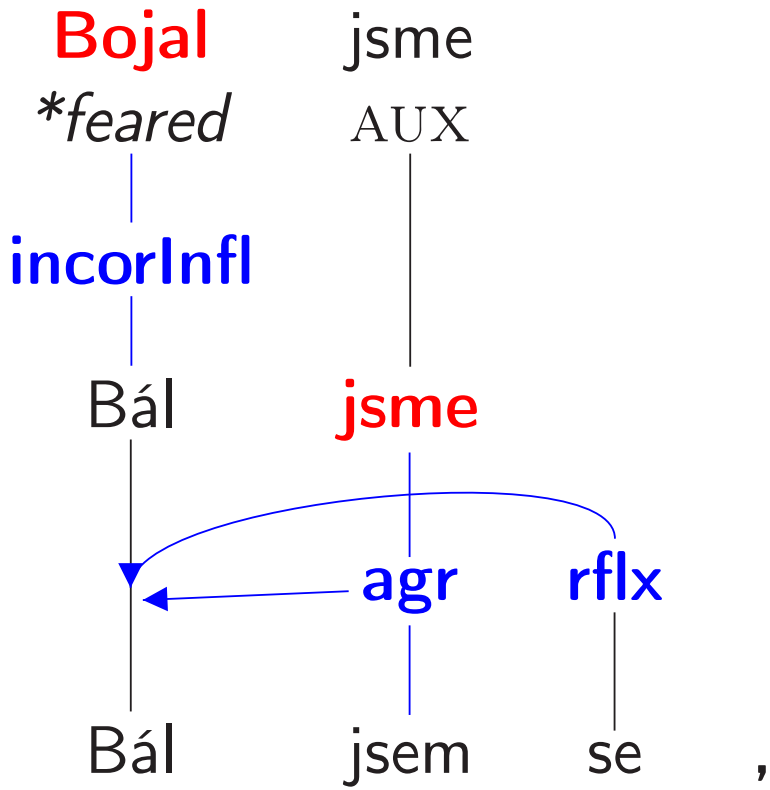
Bál jsem se, že se jí nebude líbit slavná Praha, protože to by mi velmi vadilo.

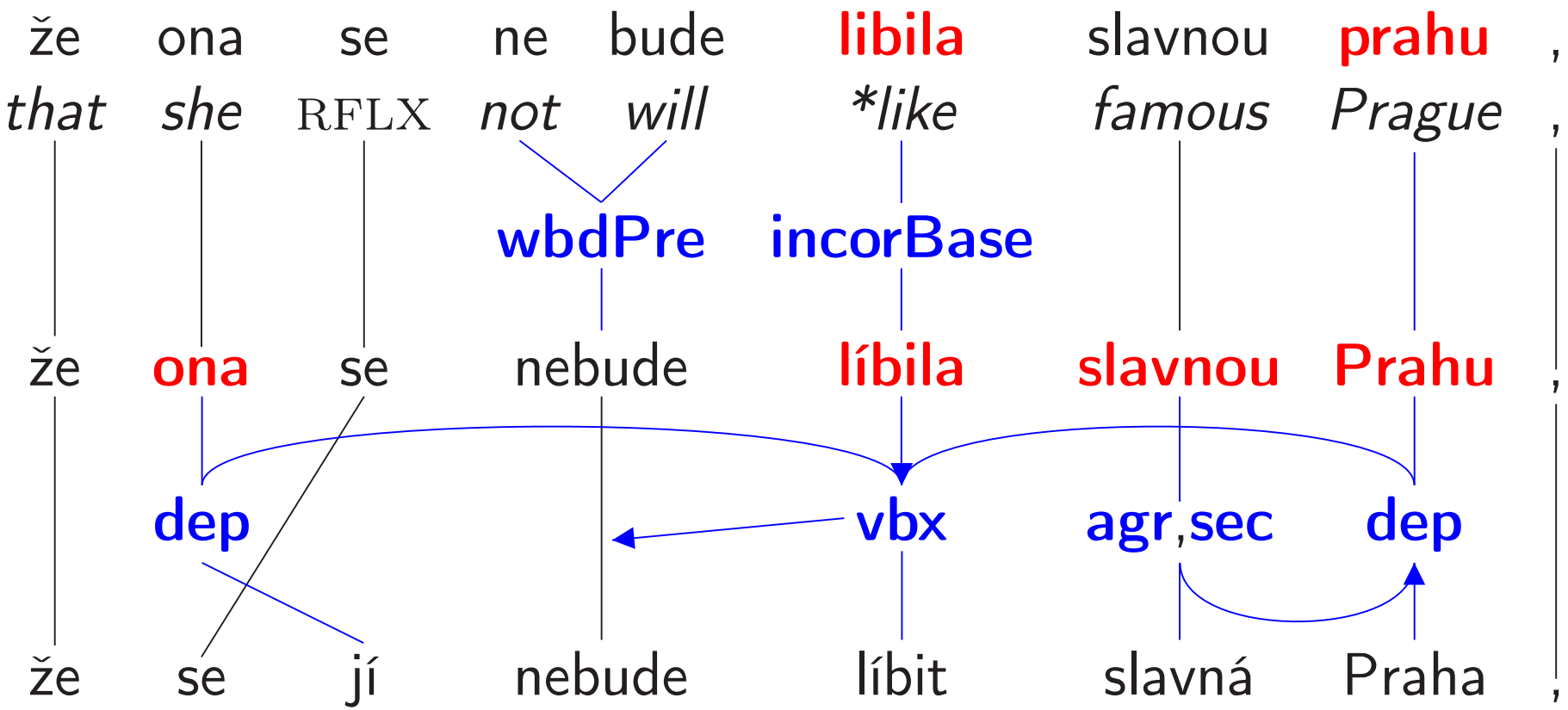
'I was affraid that she would not like the famous city of Prague, because I would be very unhappy about it.'

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

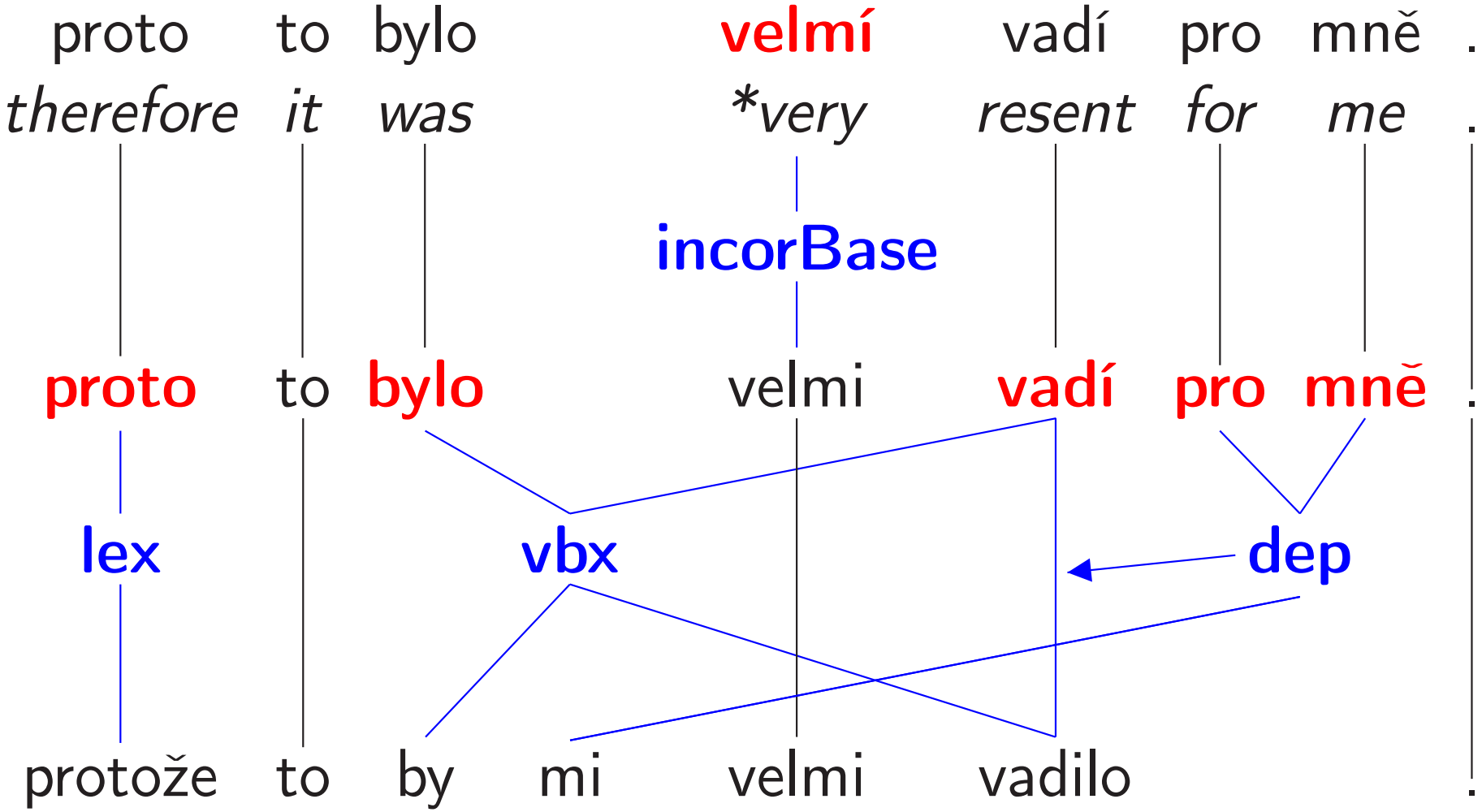
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

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

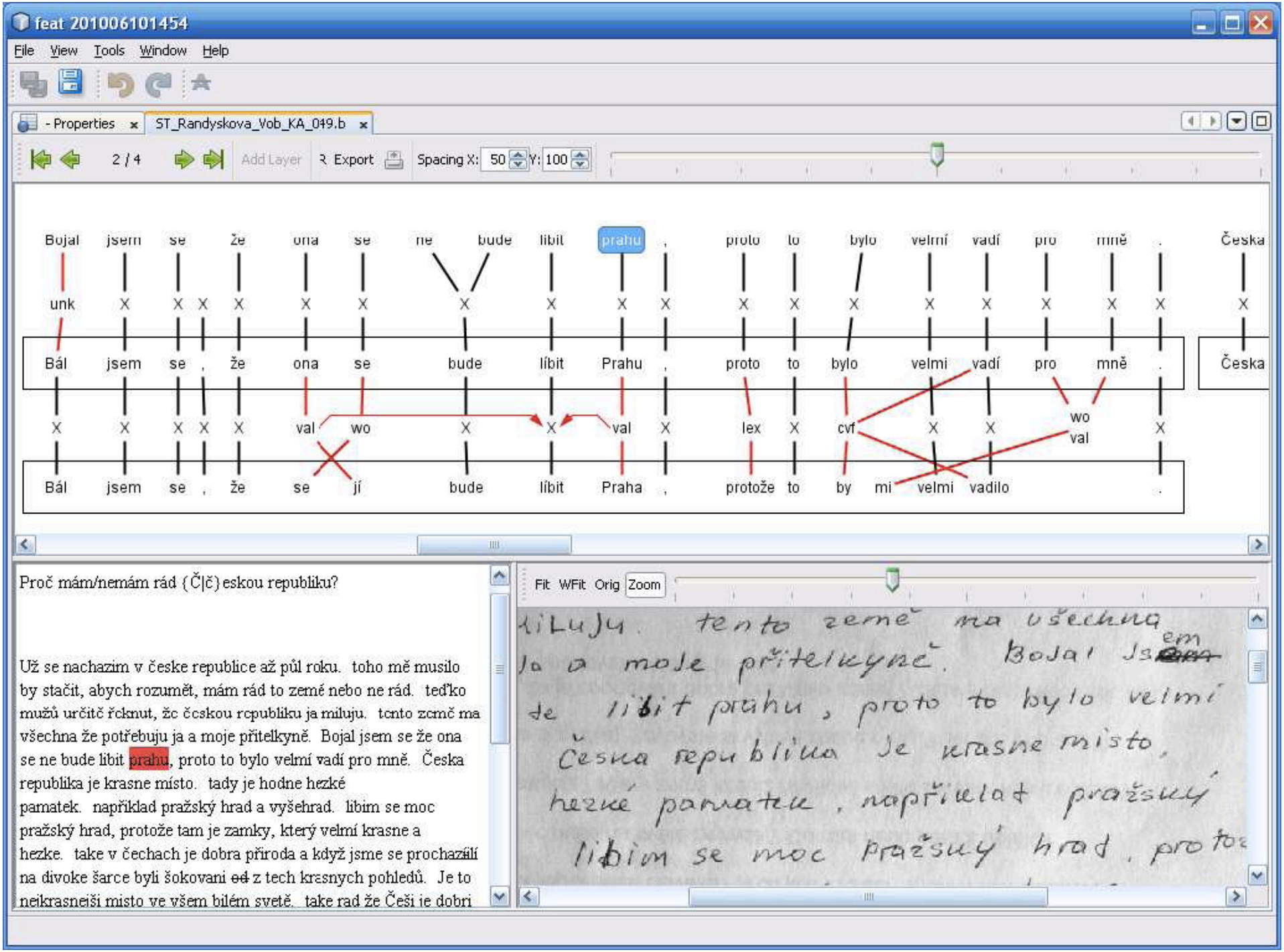

Bojal
jsme
*feared
AUX
incorInfl
Bál
jsme
agr
rflx
Bál
jsem
se
,
I was afraid



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

Osnova

- The *CzeSL-SGT* corpus:
  Czech as a **S**econd **L**anguage with **S**pelling, **G**rammar and **T**ags
- Transcriptions of essays written by non-native speakers of Czech in 2009–2013
- Extends the “foreign” part of *CzeSL-plain* by texts collected in 2013
- Transcription markup discarded
- 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[9]

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

Annotation

- If possible, each word form is tagged by a standard tagger[10] 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[11]
- The corrected text is re-tagged
- Original and corrected forms are compared and error labels, based on applicable formal criteria, are assigned[12]
- All the annotation is assigned automatically

[10] [Votrubec(2006)]
[11] [Richter(2010), Richter et al.(2012)]
[12] [Jelínek et al.(2012)]

The tools: morphological analyzer + tagger

Morče[13]

- Standard tool, reported results on native text 95–96%
- Trained on native texts (Prague Dependency Treebank)[14]
- A rule-based module deteriorates the result on learner texts
- Success varies by text

[13] [Votrubec(2006)]
[14] [PZK(2005)]

The tools: Spell-checker

Korektor[15]

- Combines rule-based morphology with a stochastic model
- Modes: spell-checker, proof-reader, diacritics assigner
- Trained on native texts (Prague Dependency Treebank)[16]
- Produces ranked suggestions with a correction type: spelling or grammar
- Suggestions for single words only, proposing single words again

[15] [Richter et al.(2012)]
[16] [PZK(2005)]

Annotation of a sample 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:		

Annotation of a sample 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:		

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*

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?
- ...

Anonymization

- The texts are anonymized by replacing personal names with appropriate forms of *Adam* and *Eva*.
- Names of smaller places (streets, villages, small towns) and other potentially sensitive data are replaced by QQQ.
- Unreadable characters or words are transcribed as XXX.

Evaluation of the automatic correction

Korektor

- The sample: 67 texts, 9373 tokens, 7995 words
- Evaluated on a manually and doubly annotated subset of CzeSL
- Using corrections where both annotators agree (97% on T1, 91% on T2)
- Ill-formed tokens:
  - total (= unknown to MA): 918
  - with identical corrections on T1: 786
- Results for ill-formed tokens:
  - diacritics assigner only: 70%
  - proof-reader: 80%
  - diacritics assigner followed by proof-reader 82%[a]

[a] [Štindlová et al.(2012)]

Osnova

Dynamic attributes

- Dynamic attributes are derived from some positions of `tag` and `tag1`.
- Useful in queries:
  - To access individual morphological categories
  - To stipulate identity of categories across mutiple forms to require grammatical concord
  - To compare values of a category for the original and corrected forms

- `k`, `k1` – word class (position 1 of the tag)
- `s`, `s1` – detailed word class (position 2 of the tag)
- `g`, `g1` – gender (position 3 of the tag)
- `n`, `n1` – number (position 4 of the tag)
- `c`, `c1` – case (position 5 of the tag)
- `p`, `p1` – person (position 8 of the tag)

Comparing annotation of original and corrected forms

Global conditions in a CQL query

- 1:[] 2:[] & 1.lemma = 2.lemma
- 1:[] 2:[] & 1.lemma = 2.word
- 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

Osnova

Theoretical issues of annotating learner language

- The absence of automatic methods and tools targeting non-native language is not caused only by the computational complexity of the task and the absence of data resources, e.g. for machine learning applications.
- There is a more fundamental issue of largely missing concepts and schemes to describe non-standard linguistic phenomena.
- An option: non-standard phenomena modelled as mismatches between different dimensions of a word class classification.[17]

[17] [Díaz-Negrillo et al.(2010), Rosen et al.(2014b), Rosen(2014)]

Word classes in 3D

Each word form has three word classes:

- **i**nflectional (morphology)
  - word as a sequence of morphs
  - to deal with ill-formed morphs or wrongly concatenated morphs
  - properties: form, lemma, paradigm
- le**x**ical (stem)
  - word as a bundle of morphemes (grammatical, lexical)
  - categories interpreted within the local context of isolated word forms
  - properties: form, lemma, paradigm, case, number, gender, person, ...
- **s**yntactic (distribution)
  - word as a syntactic constituent
  - categories interpreted in the syntactic context
  - properties: case, number, gender, person, ...

Types of mismatches:

inflectional ≠ lexical = syntactic

– to model phenomena involving morphology (stems, inflection, derivation), including distinction between a problem in stem and inflection:

- *vidím **leva*** – **iP:lev-leva**, xP:lev-lva (‘I see a lion’)
- *novoroční **předsevzení*** – iL:předsevzení, **xL:předsevzetí** (‘New Year’s resolution’)

inflectional = lexical ≠ syntactic

– to model phenomena involving morphosyntax (agreement, government):

- *vidím **lev*** – xNom, **sAcc** (‘I see a lion$_{\mathbf{NOM}}$ ’)
- *pomáhat **rodinu*** – xAcc, **sDat** (‘help the family$_{\mathbf{ACC}}$ ’)

Word classes in standard language ...

- ... can differ across dimensions, but only specific combinations are available:
- ***který*** – **iAdj,xPrn,sNoun** (‘which’)

Examples

Morphology

- *květiny* ***kvétou*** – **iP:kvést-kvétou** (‘flowers bloom’)
- *učitelka* ***bí*** *žáky* – **iP:bít-bí** (‘the teacher beats pupils’)
- *po* ***jednem*** *roku* – **iP:jeden-jednem** (‘after one year’)
- *Praha* ***libi*** *se mi moc* – **iP:libit-libi, xL:líbit** (‘Prague I like a lot’)
- ***skuzím*** – **iL:skuzit, xL:zkusit** (‘I’ll try’)
- ***mamiňkou*** – **iL:mamiňka, xL:maminka** (‘[with] maminka’)

Morphosyntactic categories

- *Chtěla bych bydlet v **nějakém evropském** zemi.* – **sFem**
  ('I'd like to live in some European$_{\mathbf{MASC}}$ country$_{\mathbf{FEM}}$')
- *na univerzitě **Karlova*** – **sLoc** (at 'Charles$_{\mathbf{NOM}}$ University')
- *skončit **magistr*** – **sAcc** ('finish Master$_{\mathbf{NOM}}$')
- *potřebuju mnoho sil a **snah*** – **xSg**
  ('need much strength and efforts')
- *nemusila jsem převzít odpovědnost za **něco*** – **sNeg**
  ('didn't have to accept responsibility for something')

Thanks to...

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

Barbora Štindlová, Jirka Hana, Tomáš Jelínek, Svatava Škodová, Karel Šebesta, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Milena Hnátková, Jan Štěpánek, Zuzanna Bedřichová, Kateřina Šormová, Kateřina Lundáková, Dagmar Toufarová

... 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)
- *PRVOUK*, the research funding programme at Charles University: *P10 – Linguistics, Acquisition and Development of Linguistic and Communicative Competence in Selected Communities of the Czech Republic*

... and you!

References I

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014).
The MERLIN corpus: Learner language and the CEFR.
In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dickinson, M. & Herring, J. (2008).
Developing online ICALL exercises for Russian.
In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*, pages 1–9, Columbus, OH.

References II


Dickinson, M. & Ragheb, M. (2009).
Dependency annotation for learner corpora.
In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories TLT8*.


Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010).
Towards interlanguage POS annotation for effective learner corpora in SLA and FLT.
Language Forum, **36**(1–2), 139–154.
Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

References III

Jelínek, T., Štindlová, B., Rosen, A., & Hana, J. (2012).
Combining manual and automatic annotation of a learner corpus.
In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue – Proceedings of the 15th International Conference TSD 2012*, number 7499 in Lecture Notes in Computer Science, pages 127–134. Springer.

Krivanek, J. & Meurers, D. (2014).
Comparing rule-based and data-driven dependency parsing of learner language.
In E. H. Kim Gerdes and L. Wanner, editors, *Dependency Theory*, Frontiers in AI and Applications. IOS Press, Amsterdam.

References IV

Meurers, D. (2013).
Natural language processing and language learning.
In C. A. Chapelle, editor, *Encyclopedia of Applied Linguistics*, pages 4193–4205. Blackwell.

Nagata, R., Whittaker, E., & Sheinman, V. (2011).
Creating a manually error-tagged and shallow-parsed learner corpus.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1210–1219, Stroudsburg, PA, USA. Association for Computational Linguistics.

References V


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


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


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.

References VI


Rosen, A. (2014).
A 3D taxonomy of word classes at work.
In L. Veselovská and M. Janebová, editors, *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, volume 4 of *Olomouc Modern Language Series*, pages 575–590, Olomouc. Palacký University.


Rosen, A., Hana, J., Štindlová, B., & Feldman, A. (2014a).
Evaluating and automating the annotation of a learner corpus.
Language Resources and Evaluation – Special Issue: Resources for language learning, **48**(1), 65–92.

References VII

Rosen, A., Štindlová, B., Škodová, S., & Hana, J. (2014b).
Using a cross-classifying taxonomy of non-standard forms to analyze non-native Czech.
In *SLE 2014 — 47th Annual Meeting of the Societas Linguistica Europaea, Workshop on Interlanguage Annotation*, Poznań, Poland. Adam Mickiewicz University.

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

References VIII

Šebesta, K. (2012).
Learner corpora and Czech language.
In I. Semrádová, editor, *Intercultural Inspirations for Language Education. Spaces for understanding.*, pages 74–89. Univerzita Hradec Králové, Hradec Králové.

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.

References IX


Štindlová, B., Škodová, S., Hana, J., & Rosen, A. (2013).
A learner corpus of Czech: current state and future directions.
In S. Granger, G. Gilquin, and F. Meunier, editors, *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve. Presses Universitaires de Louvain.