A learner corpus of Czech

Alexandr Rosen

Institute of Theoretical and Computational Linguistics
Institute of the Czech National Corpus
Charles University, Faculty of Arts

Seminarium *Przetwarzanie języka naturalnego*
Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki Polskiej Akademii Nauk
Warszawa, 21 listopad 2019

Outline of the talk

1. About learner corpora
2. Error annotation
3. Learner corpora of Czech
4. Releases of CzeSL
5. Extensions, modifications, alternatives
6. Lessons learnt

Outline of the talk

Learner Corpora

- Include texts produced by learners of a foreign language (L2)
  - also spoken
  - also native language
- Data for learners' dictionaries, later for research
  - early 1990s: *Longman Learner Corpus*
  - 2002: *International Corpus of Learner English (ICLE)*
- Used by teachers, textbook authors, FLT and SLA experts
- Error annotation: correcting and identifying non-standard forms

Using a learner corpus

To identify:

- Which error types prevail in which groups of learners
- Levels of progress in learning L2
- Optimal order and methods of teaching L2
- L1 influence
- Overuse and underuse in L2
- Features responsible for ‘foreign sound’

Annotation of Learner Corpora

Linguistic annotation

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

Error annotation

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

Outline of the talk

What is 'correct'?

Which variant of L2 as the base of comparison?

- Native speakers' prescriptive standard?
- Native speakers' standard colloquial use?
- Successful, understandable communication?
- Interlanguage – without a native standard?

Which categories?

Error types corresponding to levels in linguistic theory

- Orthography
- Morphology (morphonology, morphemics)
- Syntax (agreement, government, word order)
- Semantics/usage (reference, lexicon, aspect, tense, ...)
- Style, register

Alternatives?

- Understandability (achievement of the communication goal)
- Similarity to native speakers' texts

Error annotation

Target hypothesis

- = TH, correction, normalization, emendation, rectification
- Successive?
- Parallel alternatives or priorities?
- Follow-up?

Error tags

- = categorizing errors, error types
- No standard tagset
- The more tags the more time needed for annotation
  ASK: 23, MERLIN: 64, CzeSL: 26, RLC: 41, FRIDA: 32 + 60

Multiple error tags?

A single error can be annotated from several viewpoints

- *Přišlo hodně* ***lidi***$_{acc}$ ⇒ ***lidí***$_{gen}$ ‘Many people came’
- Spelling? Morphemics? Morphosyntax?

A single word can be ill-formed in more than one way:

- *Včera Tom* ****davovat*** ⇒ ***udělal*** *mejdan* ‘Tom threw a party yesterday’
- Spelling ⇒ *dávovat*
- Morphemics (allomorphy) ⇒ *dávat*
- Morphosyntax ⇒ *dává*
- Semantics ⇒ *dával*
- Lexeme ⇒ ***udělal***

Follow-up corrections?

- *stará se o* ***našich*** *králíkách* $\Rightarrow$ *stará se o* ***naše*** *králíky*
  'she takes care of our rabbits'
- Correction of the head only would make the context ungrammatical

Corrections as parallel text or token-related

- Depends on the language and/or error types

Outline of the talk

Merlin[1]

- Learner corpus of Czech, German, and Italian
- Matching CEFR levels with language phenomena
- Czech: 64.5k words, CEFR levels A1–C1
- Tagged, parsed, on-line searchable

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

AKCES – Acquisition corpora of Czech[2]

- An umbrella project, various funding
- Sections:
  - Native learners
  - Romani ethnolect of Czech
  - Non-native learners
- Written/spoken language

[2] [Šebesta(2012)], http://akces.ff.cuni.cz

Native learners

- Native learners, elementary and secondary school
  - Speech: *SCHOLA 2010* (1 mil. tokens, transcripts only)
  - Essays: *SKRIPT 2012* (0.7 mil. tokens)
- Romani ethnolect of Czech
  - Speech: *ROMi 1.0* (1.5 mil. token)
  - Essays: *ROMi 2013* (300 th. tokens)
- Native and ethnolect (from *SKRIPT 2012* and *ROMi 2013*
  - Essays: *SKRIPT 2015* (380 th. tokens)

Non-native learners: CzeSL – **Cze**ch as a **S**econd **L**anguage[3]

- Approx. 1 million words, transcribed hand-written essays
- L1 groups:
  - Slavic: Russian, Ukrainian, Polish, ...
  - Other Indo-European: German, English, French, ...
  - Non-Indo-European: Vietnamese, Chinese, Arabic, ...
- Searchable:
  - *KonText* (CNC) – http://kontext.korpus.cz
  - *SeLaQ* (ITCL) – http://utkl.ff.cuni.cz/czesl/selaq.html
  - *TEITOK* (ITCL) – http://utkl.ff.cuni.cz/teitok/emendace
- Downloadable:
  - http://lindat.mff.cuni.cz, https://bitbucket.org/czesl/czesl-man
  - License – Creative Commons BY-(NC-)ND 3.0

[3] http://utkl.ff.cuni.cz/learncorp/

Sizes and proportions

	whole *CzeSL*	manually annotated
Texts	8.6k	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	50–300	50–300

Number of texts by language group and proficiency level

	Slavic	IndoEuro	non-IndoEuro	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

L1	Words		Texts	
ru	572,282	57.65%	5,060	57.92%
zh	51,353	5.17%	519	5.94%
pl	49,182	4.95%	196	2.24%
uk	41,781	4.21%	382	4.37%
ko	22,560	2.27%	250	2.86%
en	21,734	2.19%	217	2.48%
de	21,722	2.19%	178	2.04%
ja	21,554	2.17%	197	2.26%
kk	21,002	2.12%	202	2.31%
vi	14,140	1.42%	121	1.39%
ar	11,778	1.19%	96	1.10%
fr	11,314	1.14%	125	1.43%
es	11,176	1.13%	132	1.51%
other	121,163	12.20%	1,061	12.15%
Total	**992,741**	**100.00%**	**8,736**	**100.00%**

Problems

- No. 1: The texts are transcribed in a standard text editor, including transcription markup, which results in many typos and inconsistences
- No. 2: Gross imbalance of L1s and profiency levels

Outline of the talk

Searchable releases[4]

- *CzeSL-plain*
  – also Romani ethnolect and native, 2 mil. words, no metadata, no annotation
- *CzeSL-SGT*
  – automatic error annotation, 1 mil. words
- *CzeSL-man v.0*
  – manually annotated, tagged, no metadata
- *CzeSL-man v.1*
  – T2 annotated with T0, tagged, parsed, errors as attributes
- *CzeSL-man v.2*
  – tagged, levels (tiers) linearized

[4] in *KonText* – http://korpus.cz except *CzeSL-man v.0*, searchable in *SeLaQ*; CzeSL-man v.1 and v.2 are available for on-line search on request

Releases published but not yet searchable

- *CzeSL-MD*
  - Includes a subset of texts from CzeSL-man, semi-automatically annotated by a multi-domain tagset with a focus on morphology,
  - Available from https://bitbucket.org/czesl/czesl-md
  - To be extended to all CzeSL-man texts
- *CzeSL-UD*
  - Texts from CzeSL-man with syntactic annotation according to the Universal Dependencies (UD) standard
  - Available from the LINDAT/CLARIN repository
- *CzeSL-man*, *CzeSL-MD* and *CzeSL-UD* will eventually be merged into a single corpus with multiple types of annotation

Outline of the talk

Hledat v korpusu

Korpus: czesl-plain v2

Typ dotazu: CQL

Vložit 'within' | Klávesnice | Předchozí dotazy

Dotaz:

```
[word="[sz]"] [word="[sz].*"]
```

V dotazu CQL můžete vložit další řádek stisknutím klávesy Sh

Výskytů: **390** | i.p.m.: **168,05** (vztaženo k celému korpusu) | ARF : **205,02** | Výsledek je promíchán

1 / 13

Výběr řádků: základní

	času . Proto každý večer	**s synem**	chodíme jenom na procházku 8.
	knihu nebo poseoivám moje rodina	**s skypi**	. Nedivám se na televizii
	Josef odsuzuje vše , co	**s sebou**	válka nese , chce se
	serem . Rád jím knedly	**s zelím**	. Nerád jím vepřove maso
	jidlo , to známena krůtu	**s zeleninami**	a zmrzlinu jako dezert .
	taky . Včera jsem mluvil	**s svým**	bratrem . Byl to dobře
	. Ale jsem se seznamila	**s sympatickym**	klukem . Moc se mi
	. On čekal na mě	**s svém**	kamarádkem , Petr . Petr
	a četla jsem krátký text	**z seriálu**	, který je přibeh o
	protože Praha je hezké město	**s starými**	budovami , hezkými uličkámí .
	zvýší . Mimochodem , jeden	**z studenti**	mluvil o Škodě . První
	, který často se hádá	**s spolužáky**	a bil je . On

	Filter	word	word	Freq	
1	p / n	sebou	s	89	
2	p / n	zahradou	s	7	
3	p / n	synem	s	6	
4	p / n	svými	s	6	
5	p / n	zajímavými	s	5	
6	p / n	svou	s	5	
7	p / n	sestřenicí	s	4	
8	p / n	sestrou	s	4	
9	p / n	zeleninou	s	4	
10	p / n	zásadou	s	3	
11	p / n	zajímavou	s	3	
12	p / n	světa	z	3	

Outline of the talk

Manual annotation of CzeSL

Ideas behind the annotation:

- Multi-layered to handle free word order and rich morphology
- Multi-purpose to cover various L1s, interpretations and users
- Error annotation grounded in L2 grammar, syntax-oriented
- Minimalist error tagset, complemented by linguistic annotation
- Minimal grammatically correct target hypothesis to improve IAA
- If in doubt, the “deeper” source of error is preferred: 1 error hypothesis per error
- Automatic assignment of some error tags to help the annotators

Multilevel Annotation Scheme

Tier 0

- Transcript of the original text

Tier 1

- Corrections disregarding word context
- Spelling, morphemics
- Result: a sequence of existing Czech forms

Tier 2

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

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

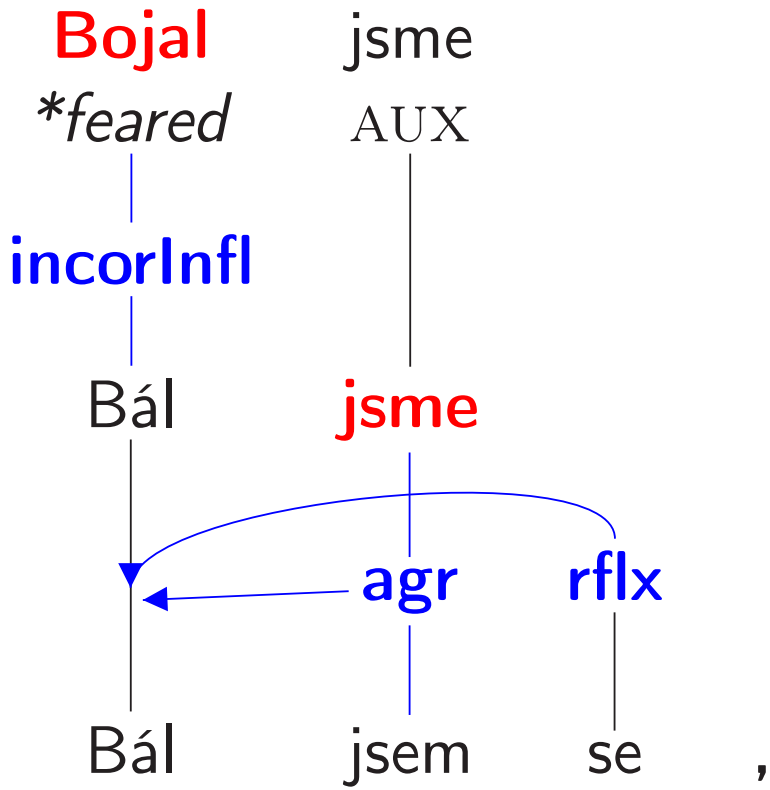
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

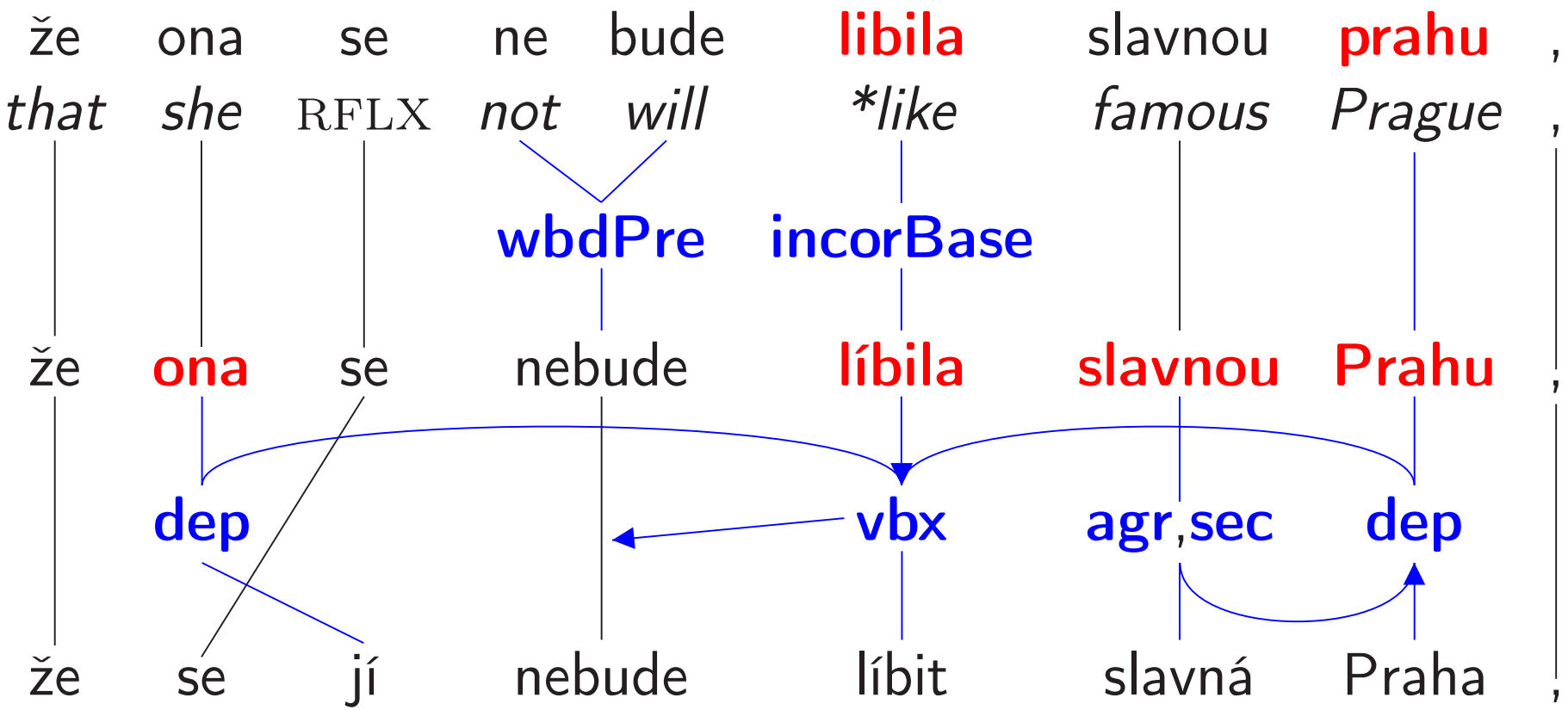
‘I was affraid that she would not like the famous city of Prague,
because I would be very unhappy about it.’

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

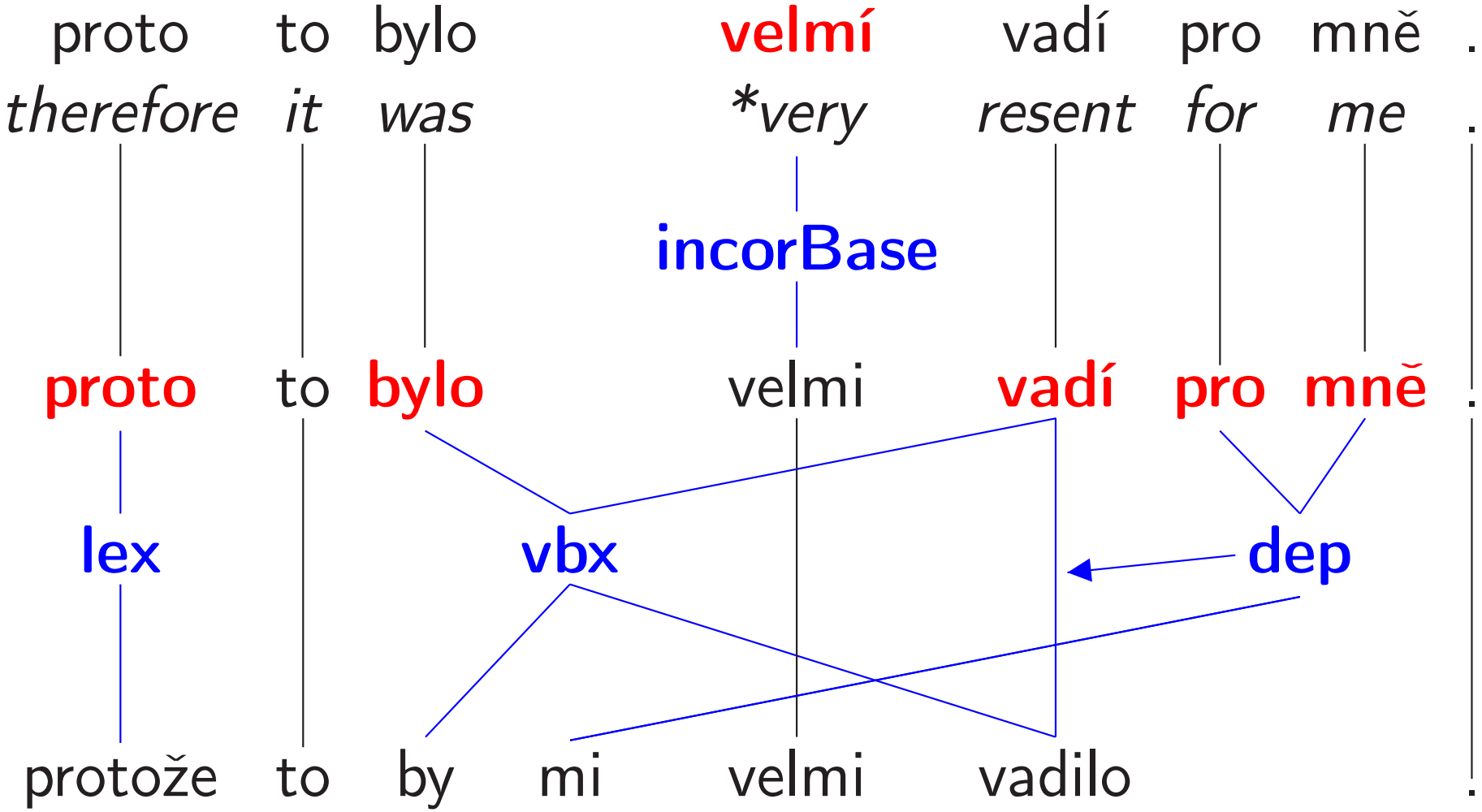
Bál jsem se, že se jí nebude líbit slavná Praha,
protože to by mi velmi vadilo.

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

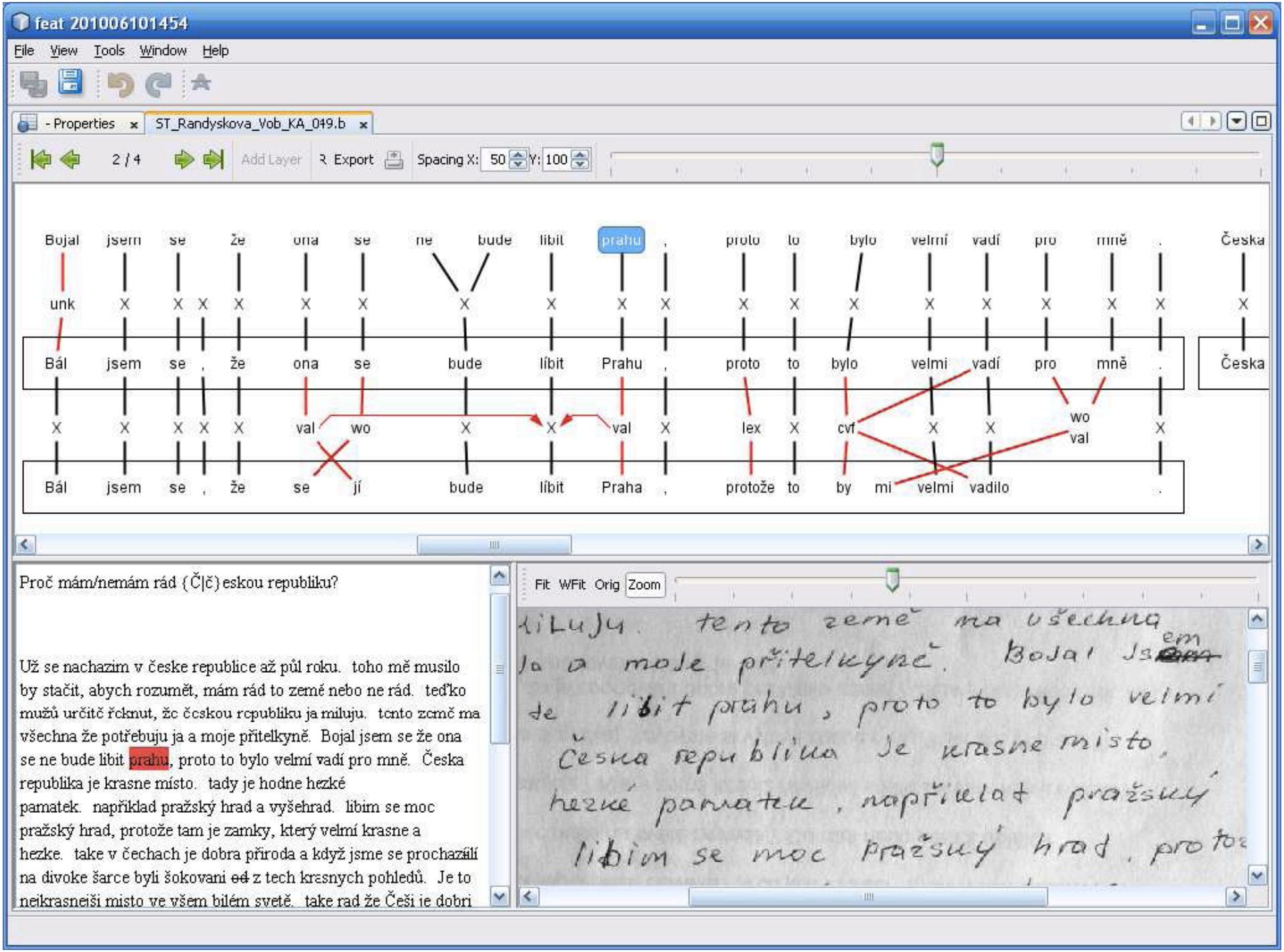

Bojal
jsme
*feared
AUX
incorInfl
Bál
jsme
agr
rflx
Bál
jsem
se
,
I was afraid



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 types at Tier 1

incor – incorrect form		
incorInfl	M	inflection error
incorBase	M	stem error
incorOther	A	other
fw – foreign word, neologism, unidentifiable		
fwFab	M	newly created “Czech” word
fwNc	M	foreign word
flex	M	inflection of fw
wbd – word-boundary error		
wbdPre	M	separate prefix, attached preposition
wbdComp	M	incorrectly separated/joined composites
wbdOther	M	other word-boundary errors

Error types at Tier 2

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

Error types at Tier 1 and Tier 2

styl – colloquial, bookish, regional expression			
	stylColl	M	colloquial expression
	stylOther	M	bookish, regional, slang expression
	stylMark	M	filler expression
problem		M	problematic phonomenon

Automatically identified error types[5]

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*

[5][Jelínek et al.(2012a)]

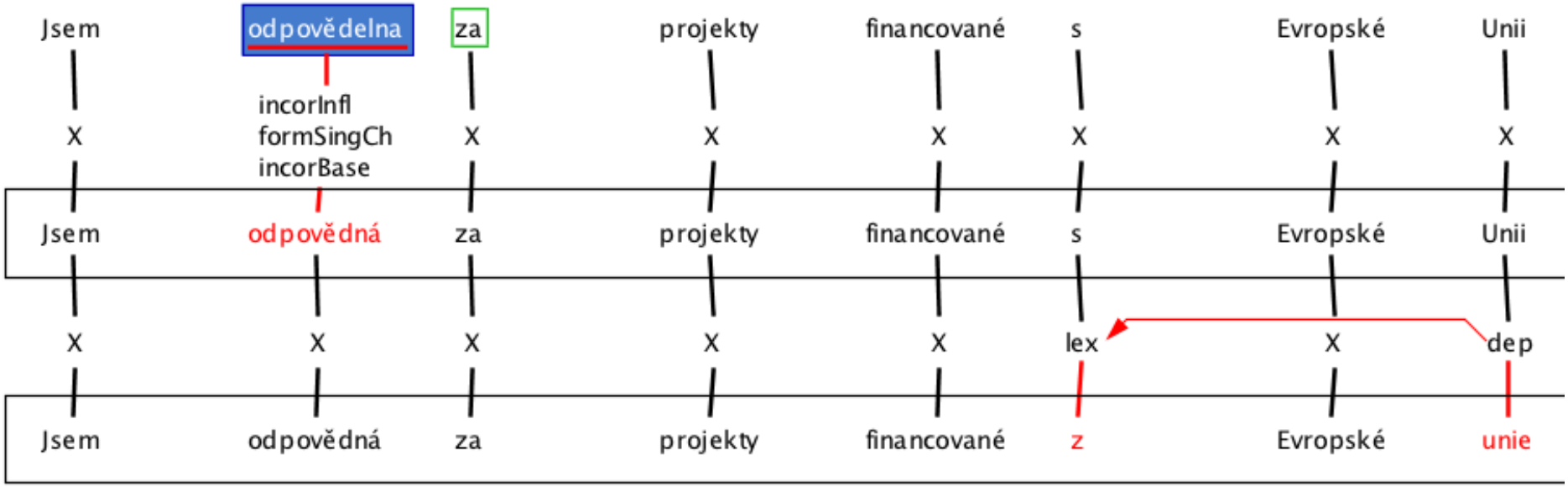
Problems

- No. 3: The error tagset is biased towards syntax
- No. 4: Some error tags have low inter-annotator agreement
- No. 5: The error annotation scheme was designed without a clear idea about a search and display tool

Annotation editor: Feat[6]

- Purpose-built
- Stand-off XML-based annotation format (*PML* – Prague Mark-up Language)
- Original: manuscript, transcription, tokenized text (Tier 0)
- Annotation:
  - spelling, morphemics (Tier 1)
  - morphosyntax, syntax, lexicon (Tier 2)

[6] https://bitbucket.org/jhana/feat/wiki/Home



Pracuju v zemský räde v Görlitz. Ale můj kancelař je v Žíta{w|vv}e. Pracuju teprve dva měsice. Studovala jsem regionalní rozvoj a pracuju v povolání. Pracuju v oddělení pro rozvoje zemský rady. Jsem odpovědelna za{přeškrtnutá čárka nad a}<co> projekty financované s Evropské Unii. Pracuju 10 hodin za tyděn.

Orig WFit Fit Zoom

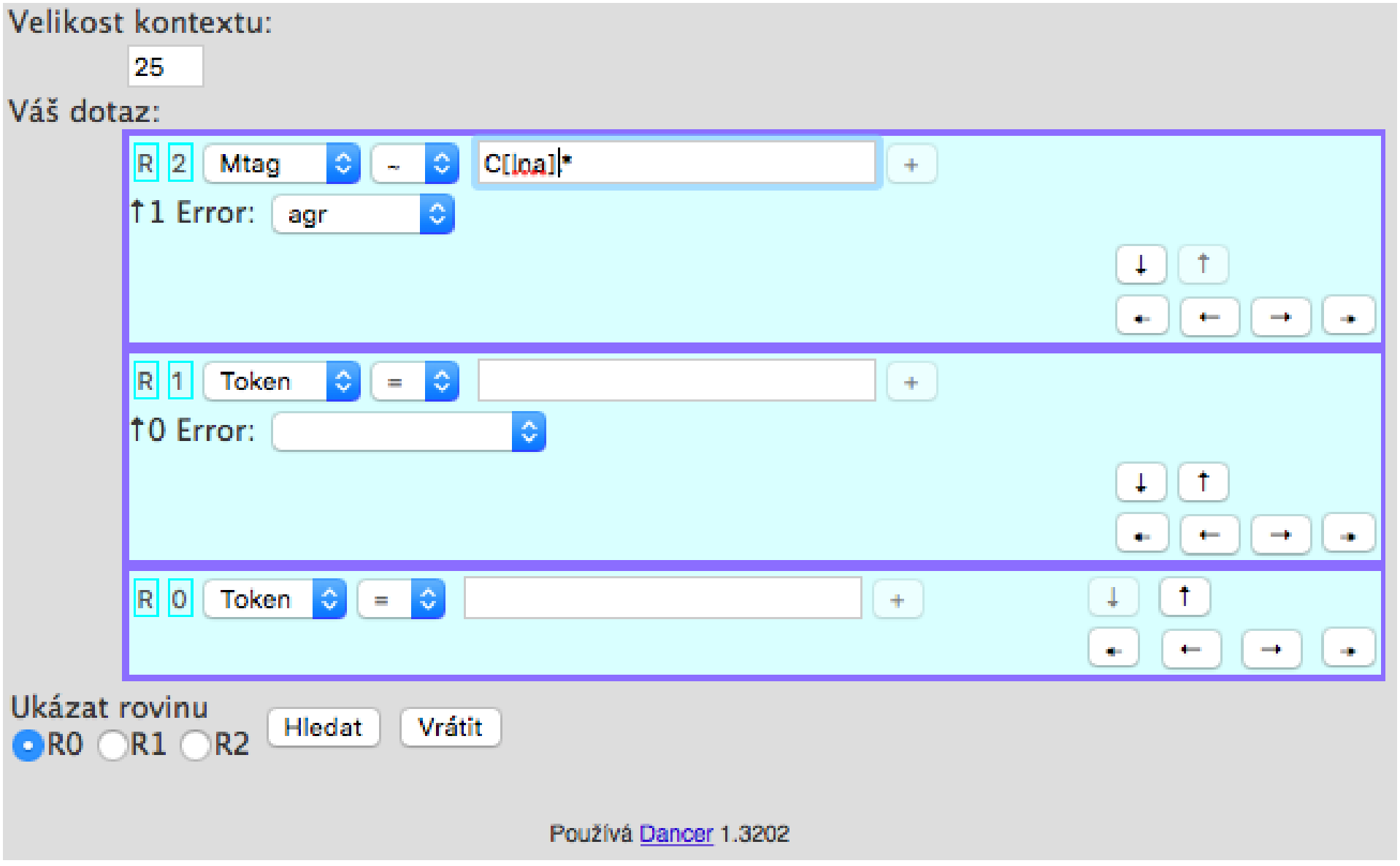
Search tool 1: SeLaQ

- Purpose-built, native in the feat format
- Queries on form, lemma, tag, error type, any level
- Shows concordances as strings of forms on any level

BUT:

- Only one tier shown at a time, no links (vertical or horizontal)
- Support for metadata missing
- No statistics

- Used for *CzeSL-man v.0* (L2 Czech+ethnolect)


Velikost kontextu:
25
Váš dotaz:
R 2 Mtag ~ C[ına]*
↑1 Error: agr
R 1 Token =
↑0 Error:
R 0 Token =
Ukázat rovinu
R0 R1 R2
Hledat
Vrátit
Používá Dancer 1.3202

ceským přitelem . Okamžitě , začala jsem studovat sáma s knihou a po **jednem**
oc hezká , měla vlasy keský , šaty červeny měla boty velkXXX a hezký a **jedna**
t všechna vejce , která mají . Když daly mu peníze a vejce , švec udělal **dva**
pět nevypadají tak strašné . Vždycky , když některy neštěstí se stane v **jednom**
. Teď studuje strojní fakulta na zapado česká univerzita bydlíme spolu **dvě**
é mi pomáhají v životě ve společenském styku . Ale teď , když už mám **jedné**
á dcera . Její dcera je taky pět let , tento rok . Takže mám dvě neteře a **jeden**
je to Adam a je jamu padesát dvit let . moje otec je dělník . moje mám **jednasetery**
chno , aby měli všecho dost . V další životě chcí mít velkou rodinu , asi **tří**
deset nebo devět roky zpátky . A samořejme ona zapoměla . Měli jsme **dvě**
aze , je to ještě těžší pro mě . Mám jen jednu sestru . Jmenuje se a je o **dvě**
chu kitara . Ona bude přiset v Praze přišti rok . Eva je moje sestra a má **dve**
ch se odpočívat , protože cesta z mého národního města do Tokia trvá **dvě**
a Praha poskytuje hodně možností : divadlo , opera , koncert . Během **dva**
jenom přežiju . Tady , můžu najit práci bez problemu protože mluvím **tři**
nocníků . Nakonec jsem se velmi bavil tím , ale bylo to i moc prácí – asi **dvě**
A ten kámen praskl zněj byl jed na opička , a ten uměl udělat kotrmec , **jeden**
epubliku , je , že mě krásný hlavní město . Praha je podle mého názoru **jedna**
udu dělat celý den v firmě . Budu mit Na oběd až 11:30 ´ v jídelně . Za **dva**
eprezentativní ani konstantní . Kulinární stránka je tahle , která spojuje **obojí**
dně . Koupili jízdenky na historickou Vídeň . Byla průvodkyně a mluvila **tři**
ky jsou hodné . Vi , co chci , co umí . Pamatovala jsem dobři . řekla mě **jednu**
n chaosu , a pak uděláme radost mamince tím , že si uklidíme . A ještě **jednó**
ůvodce mlčela , nemluvila nic . Oni našli cestou sami na konci . Stáli se **tři**
a Č|českou republiku ? Mám Českou republiku velmi ráda , proto že na **jednou**
07 obrátil v knihovně . Ráno snědl a oblékl si kabát a šli do práce . Před **čtrnáct**
t své dobře promyšlené , zajištěné blaho , nejradějí v klidu venkova . Z **jedné**
. Každý člověk musí v svém životě vybrát cíli kteři on chci dosáchnout . **Jedná**
ok) , v lete . On byl doktor a hlavně pracoval jako obvodní lékař . Mám **jeden**
. Učím se čestinu na zapadočeské univerzita . Vyberu tu skolu protože **několik**
ěno z rodinou . Aby byla obsahovala , dodála sebou jméno Eva . Až do **devět**
Ráda bych představila něco o Číně . Čína je **jedna**
, někdy divni . Ma kamaradka . najlepší . Studovali ve škole sedum let v **jednou**
mi pomáhají v životě a ve společenském styku . Ale teď , když už mám **jedně**

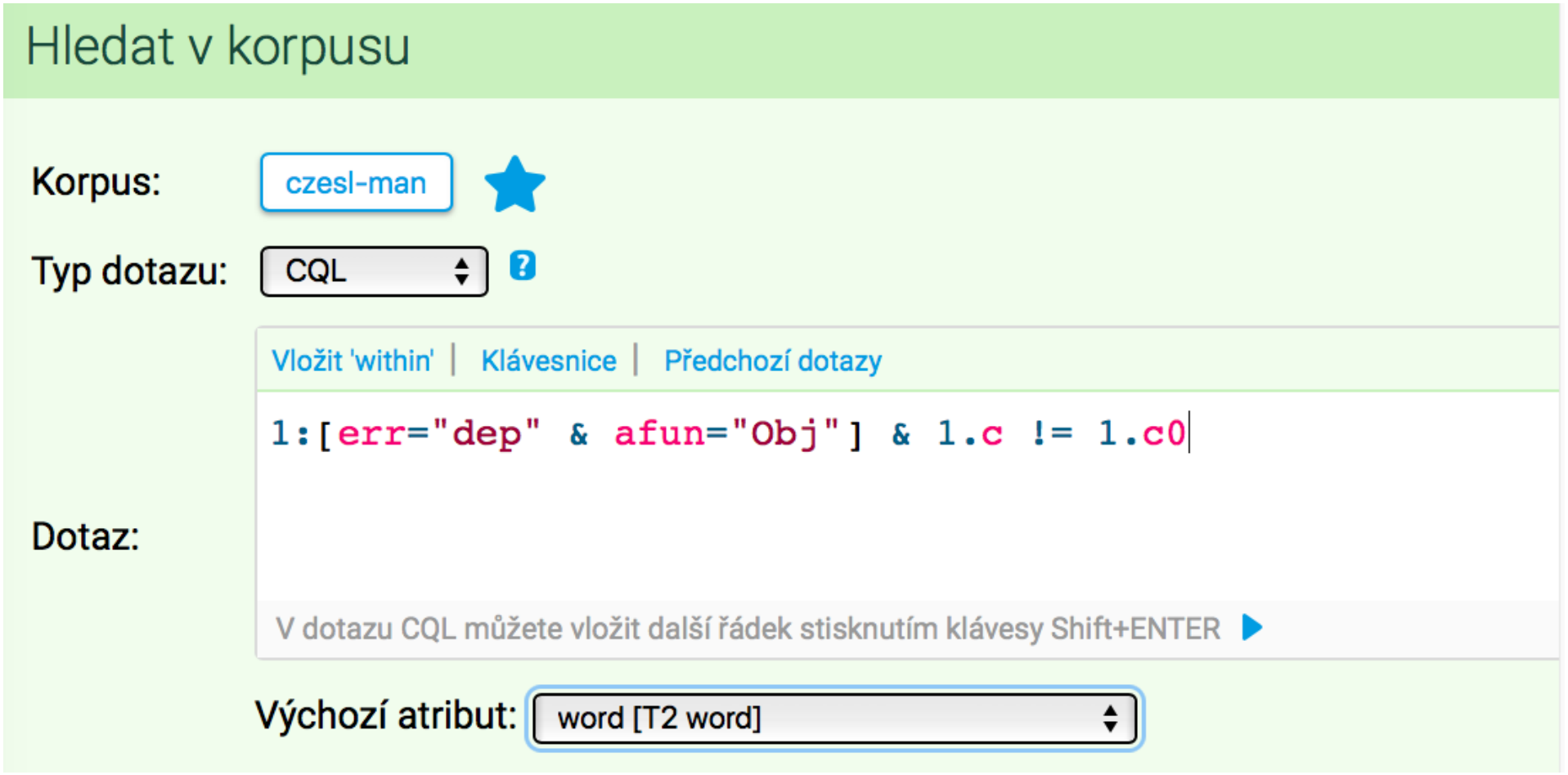
roku jsem se učastnila kurzů na českem institutu v Paržiži . Dneska , už studo
taška měla , byla sedi na laviška kouři a potom odejde koupit vice cigaretu a z
hnízda různých částech domu , položil tam vejce a přikázal pání a služce sedě
dne , myslím , že to je můj nejhorší den . Ale příští ráno se mu už směju . Tak
měsice . Ačkoli se učim česky sest měsíců , ale je samozřejmé , že ještě nemů
vysokoškolské vzdělání , chtěla bych si splnit svůj dětský sen a sen své matky
synovec . Moje máma byla zdravotní sestra ale teď nepracuje . Musí relaxovat
je to Eva a je ji dvacet tři let . moje sestry je studentka . . Adam 28 let VietNAM
, možne ctyří dětí . Myslím to bude krásne . Doufám že moji mladší sestři bud
těžké kufry , proto že jsme planovali navštěvu na měsíc . Tak máma říkala „ M
roky mladší než já . Není jen moje sestra , ale také moje nejlepší přítelkyně ze
sin . Kdy ona ma čás ona libi chodit v parku spolu jeji manžel . Eva , další sestr
dny autobusem , a nemohla jsem spát v autobusu a v letadle . Skoro tři dny js
roků , uviděla jsem víc než dese her , patnáct koncertů a tolik oper (někdy dv
jaziky a mám dobrou skušenost . Pronajmout hezký byt je taky lehké a levné .
roky jsem prováděl tého organizávání a realizace , však zatímco jsem i studov
den , byl u kamaráda a oni si hráli a potom se objevil osm zlodějů a čtyři polic
z nejzajímavějších měst v Evropě . Je tady hezká architektura , hou hodně diva
hodiny budu mit přestavka 10 minut . Budu končit pracov v 17:45 ´ a budu se
dva aspekty v celou paradoxnosti . Ačkoli vím o rodinách , které mají stejné jíc
cizí jazyky : anglicky , francouzsky a italsky . S ní jeli autobusem po městě . V
učitelku , že měla jsem lepši známky než Češi . Napřiklad , Psala jsem diktátu
zajímavost : nemám ráda pořádek , ale ráda uklízím ... Divné , že ? Je to dobře
Vídeňace .
stranu tady bydlí čast mé rodiny . Na druhou stranu miluju český jazyk . Myslí
dne se to stalo . Chtěl uznat , v kolik hodin přesně byl , tak zavolal . Opravil ča
strany taková vlastnost může být i velice prospěšná , pro mě však až na sklon
s mych takových cíl je dát moji rodičů všechno , aby měli všecho dost . V další
pes , jmenuje se Barka . On je osm let a má narozeniny v srpnu . On je hnedý a
motivů . První je můj trýc bydlí blízko Plzně . Když učím se v Plzni , můj trýc m
let , myslela jsem že , táta také měl jako středne jmeno Eva . Jaká jsem byla př
z největších států na světě . My jsme se učili ve škole , že je na třetím místě . A
třide , za jednou lavice . Ma kamáradka miluju filozofitski knihi a romaný . Cho
vysokoškolské vzdělání , chtěla bych splnit svůj dětský sen . Proto jsem přijela

Problems

- No. 6: In *CzeSL-man v.0* no metadata, no statistics, only one tier shown at a time.

Search tool 2: Manatee + KonText / Sketch Engine with error markup as token attributes

- General-purpose corpus search tool, similar to *IMS Corpus Workbench* (*CWB*)
- In-line annotation, tabular format
- Tokens (forms) and attributes (lemma, tag)
- Support for statistics and metadata
- Used for *CzeSL-man v.1*:
  - T2 is "annotated" with T0
  - T2 is tagged, lemmatized, parsed
  - T2 is assigned manual error labels
  - T0 is tagged and lemmatized


Hledat v korpusu
Korpus:
czesl-man
Typ dotazu:
CQL
Vložit 'within' | Klávesnice | Předchozí dotazy
1:[err="dep" & afun="Obj"] & 1.c != 1.c0
Dotaz:
V dotazu CQL můžete vložit další řádek stisknutím klávesy Shift+ENTER
Výchozí atribut:
word [T2 word]

Výskytů: **712** | i.p.m.: **5 755,67** (vztaženo k celému korpusu) | ARF: **380,32** | Výsledek je promíchán 1 / 24

Výběr řádků: základní

	špatného . Každému člověku ,	kdo /komu/zdát/3/1	se zdálo o zubu ,
	na Karlo-Ferdinandově univerzitě , k	čeho /čemuž/přinutit/3/2	byl přinucen otcem . Na
	císař chtěl stavět Velkou čínskou	zdi /zeď/stavět/4/3	. Vojáci si vzali hodně
	mé znalosti se neshodují se	skutečnosti /skutečností/shodovat/7/1	na 100 % . Začal
	Ukrajině , a už jsem	jich /je/vidět/4/2	neviděl dávno . Tady v
	starat , abych našel takovou	prace /práci/najít/4/-	, která by mi dávala
	se jen občas dívám na	televízí /televizi/dívat/4/-	, abych věděl , co
	říkáme , že jestli se	někdo /někomu/zdát/3/1	zdá , že druhý člověk
	v Pardubicích . Mám už	maženlka /manželku/mít/4/-	. Adam 39 let .
	do světa a navštívit všechny	zemí /země/navštívit/4/2	! Můj nejhorší den v
	naučit se češtinu , skončit	magistr /magistra/skončit/4/1	, vydělávat si peníze a
	mu moc špatně . Jen	mu /ho/bolet/4/3	bolela záda . A ještě
	pro mě . Začnu s	výživa /výživou/začít/7/1	. Vařil bych pro něj
	někdy potkal s tím neznámým	člověče /člověkem/potkat/7/5	. To znamená - lidem
	a vzpomínala na svého prvního	přítelu /přítele/vzpomínat/4/-	, jak jsme byli XXX
	řekl , že jeho teta	_ /mu/povědět/3/-	pověděla o tom , že

Problems

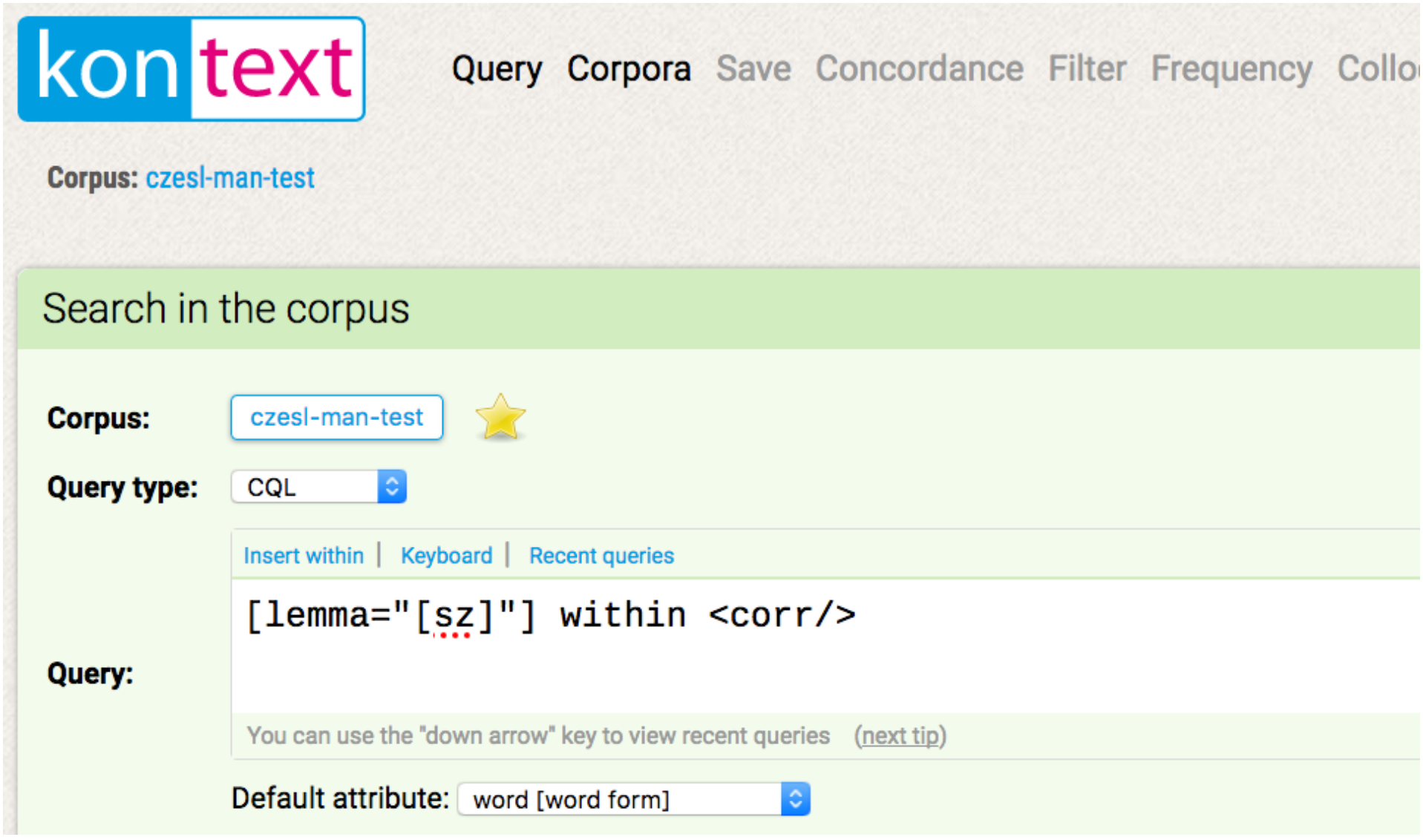
- No. 7: In *CzeSL-man v.1* only Tier 2 is preserved intact.

Search tool 2: // Manatee + KonText / Sketch Engine // with error markup as structural elements

- Structural annotation (normally used for sentences, paragraphs) can be used also for error markup: `<err/><corr/>`
- Sketch Engine offers simple learner corpus query interface
- Searching for strings in standard query interface means that queries should skip optional `err` or `corr` structures

přečístse $\Longrightarrow$ *přečíst si* ‘pročitat’ sebe’ ‘read though REFL’

```
<err tier="1" type=wbdJoin>
  přečístse
</err>
<corr tier="1" type=wbdJoin>
  přečíst
  <err tier="2" type=lex>
   se
  </err>
  <corr tier="2" type=lex>
   si
  </corr>
</corr>
```


kontext
Query
Corpora
Save
Concordance
Filter
Frequency
Collo
Corpus: czesl-man-test
Search in the corpus
Corpus:
czesl-man-test
Query type:
CQL
Insert within | Keyboard | Recent queries
Query:
[lemma="[sz]"] within <corr/>
You can use the "down arrow" key to view recent queries (next tip)
Default attribute:
word [word form]

Click to see details z, 22, pl, B1	<s>Líbí se mi , že jsem měla šanci podívat se na muslimskou zemi , která je jednou [2:dep s>z] [1:incorBase|formVoiced nejkonservativnějších>nejkonzervativnějších] , protože je to zajímavé pro mě jako arabistku .</s>
z, 22, pl, B1	<s>Vždycky [2:odd ,>] když se [2:dep mně>mě] někdo zeptá , co je pro [2:dep mně>mě] nejdůležitější v životě , [1:incorBase|formQuant0 vzpominám> [2:use vzpomínám>vzpomenu]] si na rozhovor [2:dep ze>se] známým polským filozofem [1:incorBase|formCaron1 Leszkěm>Leszkem] Kotakowskim , který jsem četla asi před dvěma lety .</s>
z, 23, pl, B1	<s>V roce 1912 se setkal s [2:dep Felici>Felicí] Bauerovou [2:use s>se] kterou byl dvakrát zasnouben v letech 1914 a 1917 v [1:incorInfl|formQuant0 te>té] době onemocněl [1:incorBase|formQuant0 tuberkulozou>tuberkulózou] .</s>
z, 23, pl, B1	<s>Potom [1:incorBase|formQuant1 prožíl>prožil] [1:incorBase|formQuant0 kratký>krátký] [1:formVoicedFin0 vztach>vztah] s [2:dep Julie>Julií] Vohryzkovou , [2:use s>se] kterou také byl zasnouben .</s>
z, 23, pl, B1	<s> [1:incorBase|formSingCh Později>Později] potkal Milenu Jesenskou , [2:use s>se] kterou se viděl jen [1:incorBase|formQuant0 parkrát>párkrát] , ale dlouhou dobu udržovali korespondenční styk .</s>
z, 23, pl, B1	<s>Chmurnost jeho příběhů vyplývá [2:use|sec z>ze] [2:lex ukončení>zakončení] , která jsou zpravidla [2:agr tragické>tragická] , bezvýchodná , nepotěšující .</s>
z, 21, pl, B2	<s>Líbí se mi , že [1:incorBase|formMissChar potřeuju>potřebuju] jenom [2:dep jednou>jednu] kartu – studentský průkaz ISIC a že všechno můžu [1:formVoicedFin1 z>s] [2:dep ni>ní] udělat .</s>
z, 25, pl, A2+	<s>Jsem [1:incorInfl|incorBase|formSingCh odpovědelna>odpovědná] za projekty financované [2:lex s>z] Evropské [2:dep Unii>unie] .</s>

Problems

- No. 8: In *CzeSL-man v.2* some error types cannot be shown.

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[7]

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

Annotation

- Original is tagged and lemmatized[8]
- Incorrect forms are corrected by a spelling/grammar checker[9] (82% success)
- The corrected text is re-tagged
- Original and corrected forms are compared and formal error labels are assigned[10]

[8] [Votrubec(2006)]
[9] [Richter(2010), Richter et al.(2012)]
[10] [Jelínek et al.(2012b)]

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:		

Hledat v korpusu

Korpus: czesl-sgt / --celý korpus--

Typ dotazu: CQL

Dotaz:

Vložit tag | Vložit 'within' | Klávesnice | Předchozí dotazy

```
1:[tag!="X@.*" & tag1="N.*"] & 1.c != 1.c1
```

V případě problémů s kompatibilitou pokročilého CQL editoru s vaším prohlížečem m
základního CQL editoru v menu "Zobrazení" → "Obecné volby zobrazení"

Výskytů: **403** | i.p.m.: **351,21** (vztaženo k celému korpusu) | ARF: **207,94** | Výsledek je promíchán 1 / 14

Výběr řádků: základní

☐	ja ✦ B1	jsem šla na procházku v	Paříže /Paříži/SingCh	. Šla jsem do kavárny
☐	zh ✦ A1	kamarádku . Ju . Má	černa /černá/Quant0	vlasy . Má rada dancovat
☐	zh ✦ A2	drahý . měj se hezký	Sin /sen/Cap1	V současné době Facebook je
☐	ja ✦ A2+	. Ale tam byl šest	lidé /lidí/SingCh	. Do té doby jsem
☐	zh ✦ A2+	. Ale jsem neměla dobrý	napad /nápad/Quant0	, a jenom jsem cvičila
☐	vi ✦ A2	den a potom poideme do	prahy /Prahy/Cap0	, proteře jsem slyšel že
☐	ko ✦ A1	se nebude spočitát v mem	uzemí /území/Quant0	. Taky aby cizinci byly
☐	ko ✦ A1	půl jednácteho sis divala na	televizí /televizi/Quant1	. Dívala sis na televizí
☐	ja ✦ B1	Nejdřív jsme šli do realitní	kanceláři /kanceláře/SingCh	. Chtěli jsme zařízený pokoj
☐	vi ✦ A2+	, na horý nebo do	prahy /Prahy/Cap0	asi na jeden týden .
☐	ko ✦ A2	. Kdy budeš jet do	prahy /Prahy/Cap0	? Těším se na tebe
☐	ja ✦ A2+	vytvořen lidmi , kteří jsou	říkáni /říkání/Quant0	, že oni jsou předek
☐	zh ✦ A1	se 7 hodin večer v	Olomouce /Olomouci/SingCh	restauraci . Těším se ,
☐	ja ✦ A2+	. Taky šli do cestovní	kanceláři /kanceláře/SingCh	. Byly nábidky historické ,
☐	zh ✦ B1	média už mění styl našeho	životu /života/SingCh	. Už jsme si zvykli
☐	zh ✦ A2	dály ještě vino . To	vino /víno/Quant0	šumavy . Táky nám chut
☐	zh ✦ A2	to bude těšké . V	poděli /pondělí/SingCh	18:00 se budu dívat na
☐	ko ✦ A2	hory je nejkrásnější hory v	Koreje /Koreji/SingCh	. A potom lyžujeme a
☐	ko ✦ A2	často miminko umřel před 100	dni /dny/Y0	. Proto jsme oslavovali 100

Problems

- No. 9: In *CzeSL-SGT* the automatic correction is not reliable

Outline of the talk

Problems summary

- No. 1: The texts are transcribed in a standard text editor, including transcription markup, which results in many typos and inconsistences
- No. 2: Gross imbalance of text types in the corpus
- No. 3: The error tagset is biased towards syntax
- No. 4: Some error tags have low inter-annotator agreement
- No. 5: The error annotation scheme was designed without a clear idea about a search and display tool
- No. 6: In *CzeSL-man v.0* no metadata, no statistics, only one tier shown at a time
- No. 7: In *CzeSL-man v.1* only Tier 2 is preserved intact
- No. 8: In *CzeSL-man v.2* some error types cannot be shown.
- No. 9: In *CzeSL-SGT* the automatic correction is not reliable

Error tagging alternatives

- Implicit error annotation derived from correction?
  - COPLE2: orthographic, morphosyntacti, lexical errors[11]
- Errors as mismatches between linguistic levels?
  - Morphology vs. syntax vs. lexicon[12]

[11] [Mendes et al.(2016)]
[12] [Díaz-Negrillo et al.(2010)]

The MD error tagset

- Complements the old tagset
- Focus on morphology
- A single error may be annotated by more error types in parallel

TEITOK – a tool for creating, editing and viewing corpora

- TEITOK = TEI + TOKenization[13]
  - TEI: rich textual mark-up, according to standards for digitized texts (Text Incoding Iniciative[14])
  - TOK: linguistic annotation of tokenized text
- Web-based GUI[15]
- Different visualization and editing modes, including searchable facsimile images
- Efficient search via Corpus Workbench, using Corpus Query Language
- Multiple corpora searchable

[13] [Janssen(2016)]
[14] http://www.tei-c.org
[15] http://teitok.corpuswiki.org

- http://utkl.ff.cuni.cz/teitok/czesl/index.php?action=file&id=TESTS/NEM_GD_008.xml

TESTS/NEM_GD_008.xml

Without Title

view teiHeader

View options

Text: Transcription | Written form | Normalized form – Show: Colors | Formatting | <pb> | <lb> | Images – Tags: POS tag | UD POS tag | UD features | Lemma

Edit the information about each word of this file by clicking on the word in the text below, or click here to edit the raw XML

Bratr a Sestra.

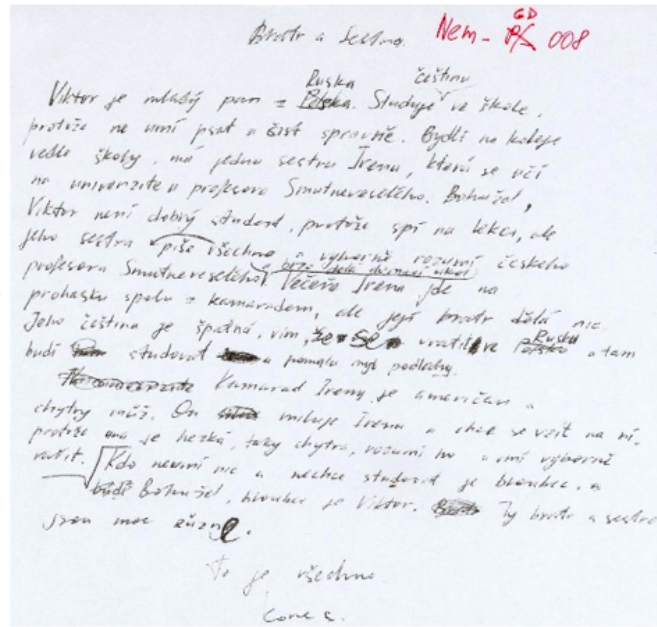
Viktor je mladý pan z ~~Polska~~Ruska. Studuje češtinu ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra ~~piše všechno~~ všechno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

Kamarad Ireny je američan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

Kdo neumí nic a nechce studovat je bloubec. ~~budi~~ Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~y~~e.

To je všechno.

Konec



Raw Verticalized Corpus View

XML File: TESTS/NEM_GD_008.xml

<tok id="w-1" deprel="root" upos="NOUN" xpos="NNMS1-----A----" feats="Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Pos" lemma="bratr">Bratr</tok>

<tok id="w-2" deprel="cc" upos="CCONJ" xpos="J^----------" feats="_" lemma="a" head="w-3">a</tok>

<tok id="w-3" deprel="conj" upos="NOUN" xpos="NNFS1-----A----" feats="Case=Nom|Gender=Fem|Number=Sing|Polarity=Pos" lemma="sestra" head="w-1">Sestra</tok>

<tok id="w-4" deprel="punct" upos="PUNCT" xpos="Z:----------" feats="_" lemma="." head="w-1">.</tok>

<tok id="w-5" deprel="nsubj" upos="PROPN" xpos="NNMS1-----A----" feats="Animacy=Anim|Case=Nom|Gender=Masc|NameType=Giv|Number=Sing|Polarity=Pos" lemma="Viktor" head="w-8">Viktor</tok>

<tok id="w-6" deprel="cop" upos="AUX" xpos="VB-S---3P-AA---" feats="Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act" lemma="bÃ½t" head="w-8">je</tok>

<tok id="w-7" deprel="amod" upos="ADJ" xpos="AAMS1----1A----" feats="Animacy=Anim|Case=Nom|Degree=Pos|Gender=Masc|Number=Sing|Polarity=Pos" lemma="mladÃ½" head="w-8">mladÃ½</tok>

<tok id="w-8" deprel="root" upos="NOUN" xpos="NNMS1-----A----" feats="Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Pos" lemma="pan" nform="pÃ¡n">pan</tok>

<tok id="w-9" deprel="case" upos="ADP" xpos="RR--2----------" feats="AdpType=Prep|Case=Gen" lemma="z" head="w-10">z</tok>

<tok form="Ruska" id="w-10" deprel="nmod" upos="PROPN" xpos="NNNS2-----A----" feats="Case=Gen|Gender=Neut|NameType=Geo|Number=Sing|Polarity=Pos" lemma="Rusko" head="w-8">~~Polska~~<add>Ruska</add></tok>

<tok id="w-11" deprel="punct" upos="PUNCT" xpos="Z:----------" feats="_" lemma="." head="w-8">.</tok>

<tok id="w-12" deprel="root" upos="VERB" xpos="VB-S---3P-AA---" feats="Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin|Voice=Act" lemma="studovat">Studuje</tok>

<tok form="Ä�eÅ¡tinu" id="w-13" deprel="obj" upos="NOUN" xpos="NNFS4-----A----" feats="Case=Acc|Gender=Fem|Number=Sing|Polarity=Pos" lemma="Ä�eÅ¡tina" head="w-12">Ä�eÅ¡tinu</tok>

<tok id="w-14" deprel="case" upos="ADP" xpos="RV--6----------" feats="AdpType=Voc|Case=Loc" lemma="v" head="w-15">ve</tok>

<tok id="w-15" deprel="nmod" upos="NOUN" xpos="NNFS6-----A----" feats="Case=Loc|Gender=Fem|Number=Sing|Polarity=Pos" lemma="Å¡kola" head="w-13">Å¡kole</tok>

Verticalized Corpus View

XML File: TESTS/NEM_GD_008.xml

	Transcription	Normalized form	UD POS tag	Lemma
w-1	Bratr		NOUN	bratr
w-2	a		CCONJ	a
w-3	Sestra		NOUN	sestra
w-4	.		PUNCT	.
w-5	Viktor		PROPN	Viktor
w-6	je		AUX	být
w-7	mladý		ADJ	mladý
w-8	pan	pán	NOUN	pan
w-9	z		ADP	z
w-10	~~Polska~~Ruska		PROPN	Rusko
w-11	.		PUNCT	.
w-12	Studuje		VERB	studovat
w-13	češtinu		NOUN	čeština
w-14	ve		ADP	v
w-15	škole		NOUN	škola
w-16	,		PUNCT	,
w-17	protože		SCONJ	protože
w-18	ne umí	neumí	PART+VERB	neumět
w-19	psat	psát	VERB	psat
w-20	a		CCONJ	a
w-21	číst	číst	VERB	sečin
w-22	spravně	správně	ADV	spravně
w-23	.		PUNCT	.
w-24	Bydlí		VERB	bydlet
w-25	na		ADP	na
w-26	koleje	koleji	NOUN	kolej
w-27	vedle		ADP	vedle
w-28	školy		NOUN	škola

Error Annotation

Without Title

Annotation of phrasal errors using the error code system of the Cambridge Learner Corpus, adopted for Portuguese errors.

Bratr a Sestra.

Viktor je mladý pan z Ruska. Studuje češtinu 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, Viktornení dobrý student, protože spí na lekci, ale jeho sestra všechno piše a vyborně rozumí českeho profesoraSmutneveseleho a brzo delá domací ukol. Večeře Irena jde na prohaskuspolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve Rusku a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je američan a chytry můž. On miluje Irenu a chce se vzít na ní. protoženo je hezká, taky chytra, rozumí ho a umí vyborně vařit.

Kdo neumí nic a nechce studovat je bloubec. budi Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různe.

To je všechno.

Konec

text view

Edit Annotation

Selection:
kamaradem

Type	[select]
Code	
Correction	

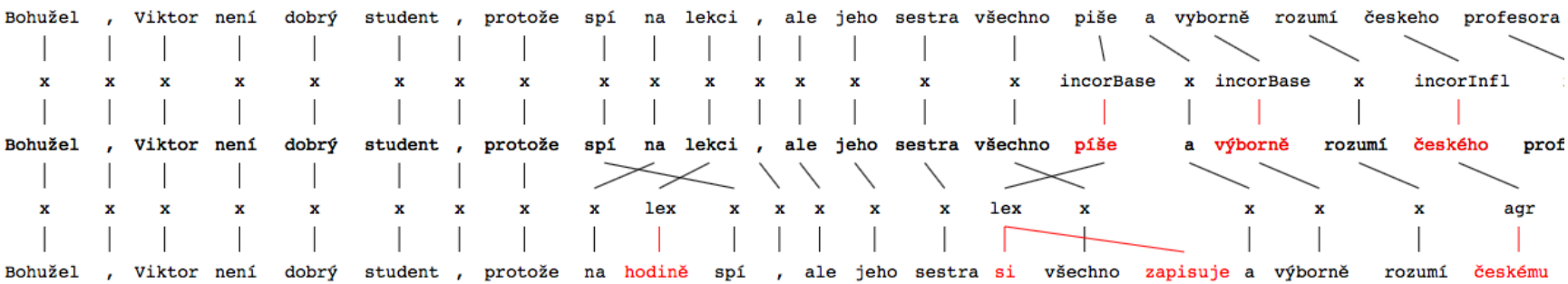
Save Cancel

NEM_GD_008

Feat

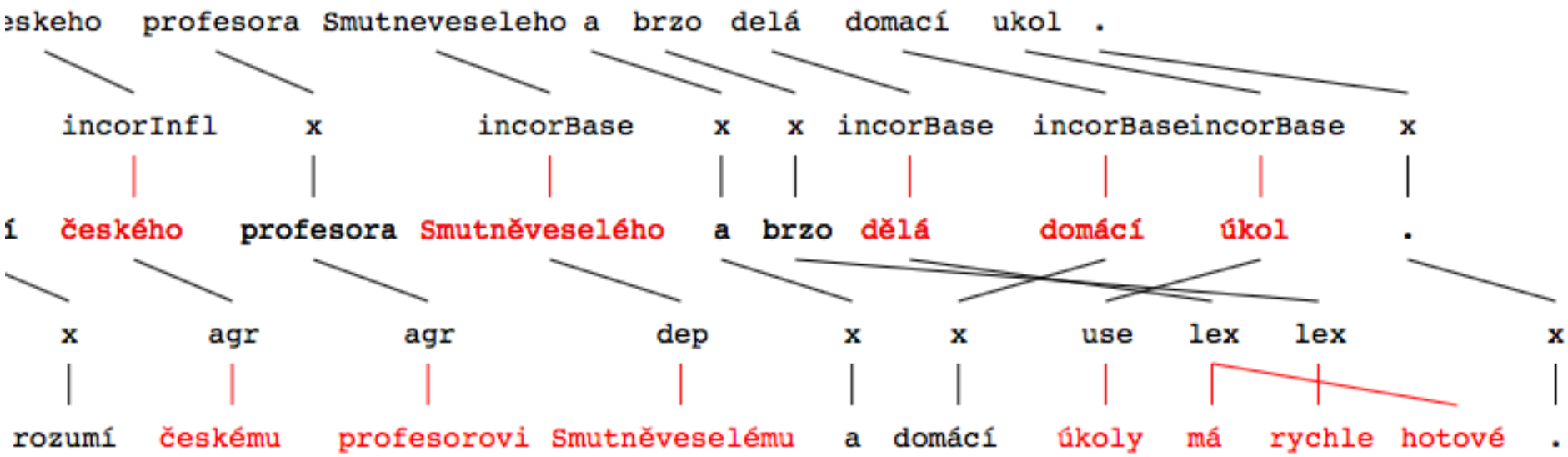
Sentence: d1p2s5

Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra ~~piše všechno~~ všechno piše a vyborně rozumí českeho profes



Sentence list · Text view · Download SVG · Download PNG

českeho profesora Smutneveseleho a brzo delá domací ukol.



Outline of the talk

Planning

- Don't start collecting data before you clearly see the end of it all.

Choice of texts

- Texts should be collected in a way to give a balanced corpus (L1 and CEFR).

Transcription

- Should be done in an XML editor, checking for transcription marks on-line.

Tools

- A compatible and ready-to-use toolchain, including annotation task manager, should be available as early as possible.
- Search tool!

Thanks to...

Jirka Hana
Maarten Janssen
Tomáš Jelínek
Barbora Štindlová
Svatava Škodová
Vojtěch Kovář
Pavel Procházka
Hana Skoumalová
...

... 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)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

References II


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.


Janssen, M. (2016).
Teitok: Text-faithful annotated corpora.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

References III


Jelínek, T., Štindlová, B., Rosen, A., & Hana, J. (2012a).
Combining manual and automatic annotation of a learner corpus.
In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, pages 127–134. Springer Berlin Heidelberg, Berlin, Heidelberg.


Jelínek, T., Petkevič, V., Rosen, A., & Skoumalová, H. (2012b).
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.

References IV


Mendes, A., Antunes, S., Janssen, M., & Gonçalves, A. (2016).
The COPLE2 corpus: a learner corpus for Portuguese.
In N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).


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

References V


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.

References VI


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.