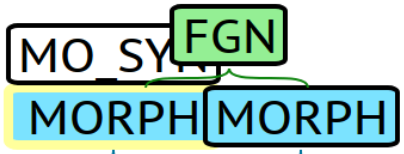
New error annotation of Czech learner corpora

Tomáš Jelínek,
Faculty of Arts, Charles University, Prague

Tím způsobem jsme musely být celou dobu v San'é , a na výlety jsme jezdily jenom s jinými studenty k Rudému moři .

Navzdory tomu se mi San'á moc líbila , protože má krásné Staré Město s výjimečnou architekturou , jakou nikde v Evropě nenajdeme .



Kromě toho má San'á úžasný klimat , protože leží ve výšce 2 200 m .

Outline

- **Motivation**
- **CzeSL corpus**
- **Annotating errors by levels of language description**
- **Brat annotation tool**
- **Automatic text preprocessing**

Motivation

Hodně **lidi** si myslí, že literatura je nudná.

Hodně **lidí** si myslí, že literatura je nudná.

Many **people**nom.pl. think that literature is boring.

- Missing diacritic?
- Wrong vowel length?
- Wrong form for gen.pl.?
- Nom.pl. instead of gen.pl. (subject / quantified noun)?

Motivation

Mám štěstí , že mám dobré **kamarádi**,

Mám štěstí , že mám dobré **kamarády**,

I'm lucky that I have good **friends**$_{\text{nom.pl.}}$

- Wrong orthography i/y?
- Wrong consonant (d/ď)?
- Wrong form for accusative pl.?
- Nom.pl. instead of accusative pl.?

Motivation

napsat magisterskou **prace**

napsat magisterskou **práci**

to write master's **thesis**

Čína je **jedna** z největších států NOT **jed_na**

Čína je **jeden** z největších států NOT **jeden**

China is **one** of the biggest states

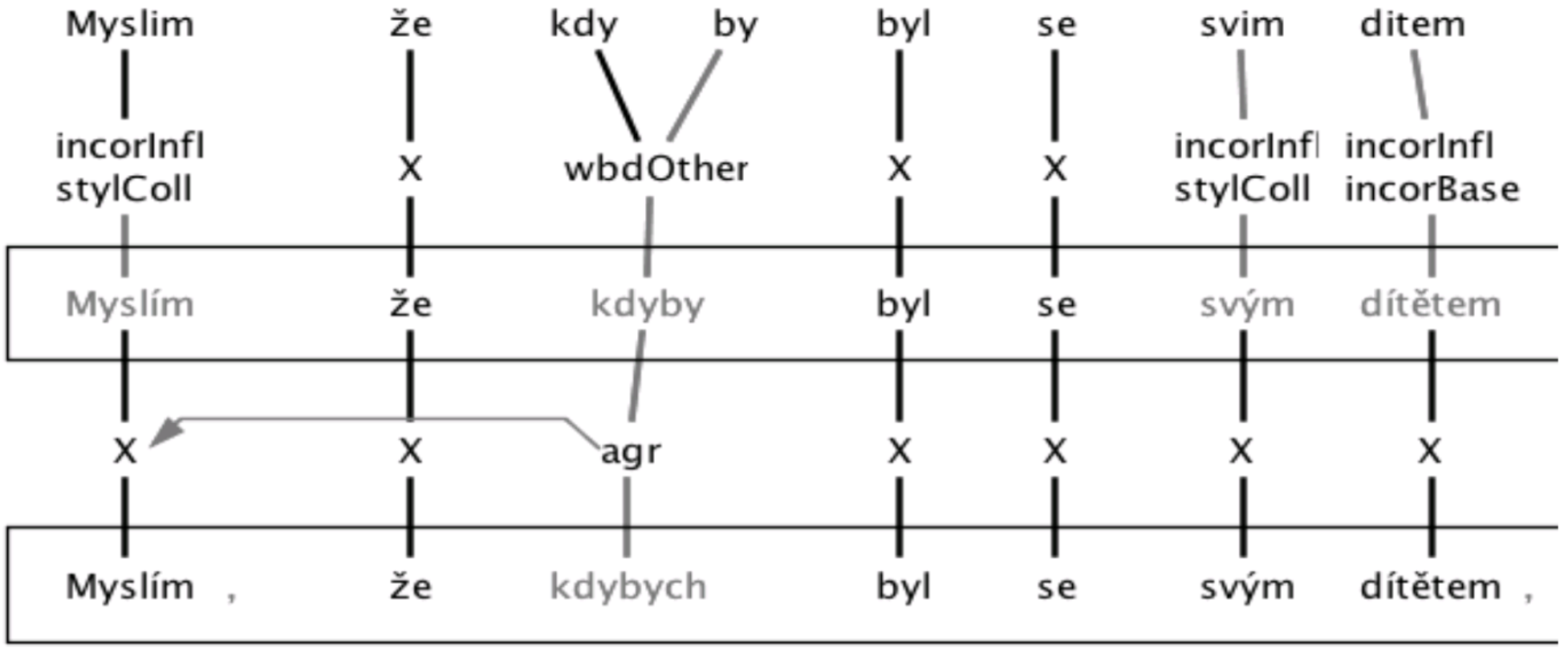
Motivation

magisterskou **prac | e**

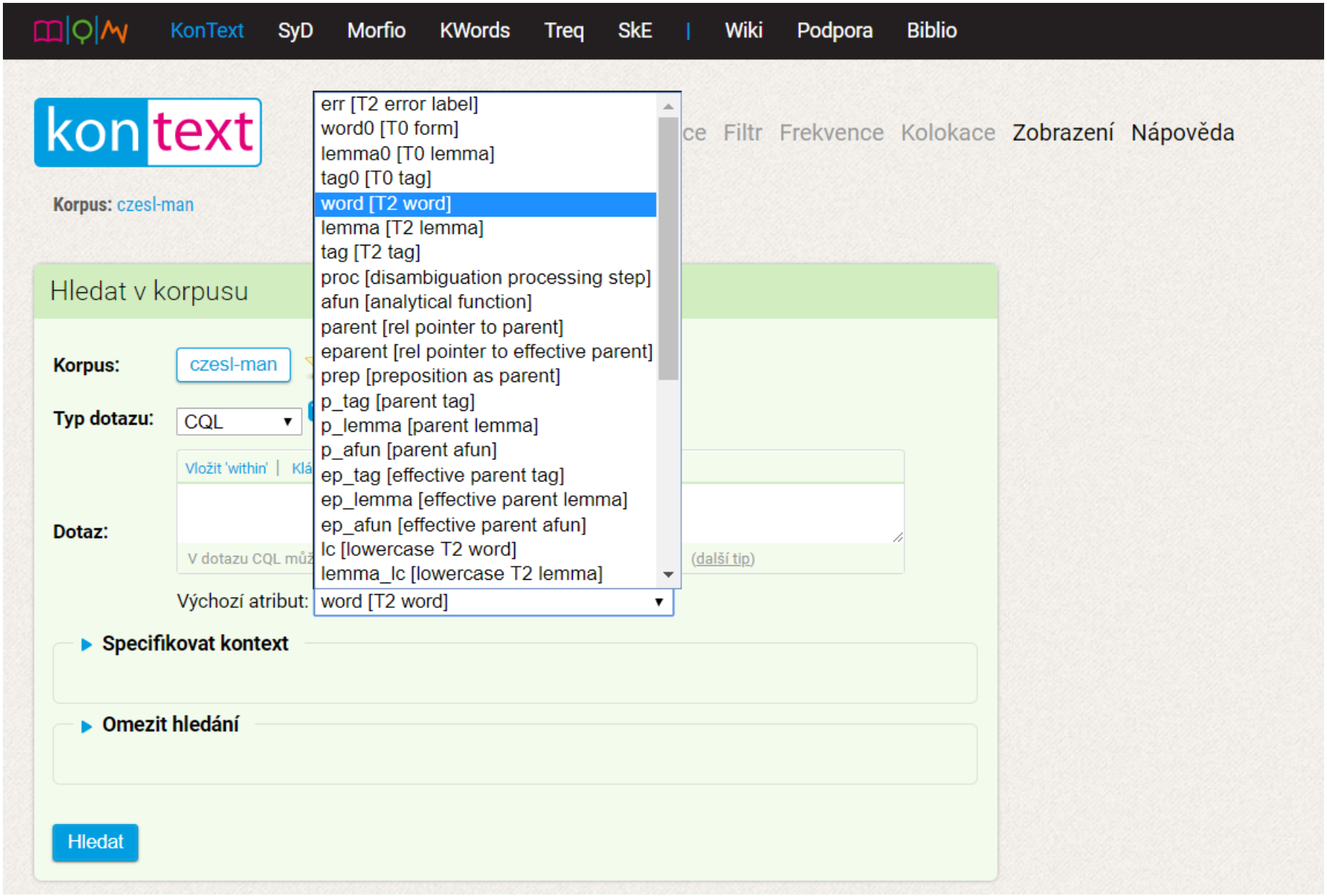
magisterskou **práci**

master’s **thesis**

CzeSL corpus



CzeSL corpus



Error annotation by levels of linguistic description

Domains

Orthography **ORT**

Morphonology and phonology **MPHON**

Morphology **MORPH**

Syntax **SYNT**

Lexicon and use **LEX**

… SEC, PROBL

Orthography

CAP **capitalization** *praha / Praha*

SEG **segmentation** *ne bral / nebral*

PUN **punctuation** *- / ,*

GEM **gemination** *nejednodušší / nejjednodušší*

SUBST **char. subst. without pron. change** *yako/jako*

- ***IY*** **i/y** *visoká / vysoká*
- **MNE** **mě/mně** *mněsto / město*
- **U** **ú/ů** *múže / může*

Morphonology and Phonology

MET	**metathesis**	*jesm / jsem*
VOC	**vocalisation**	***s** / **se** svojí rodinou*
EP	**epenthesis**	*o**db**ereme / o**deb**ereme*
SOFT	**soft cons.**	*šť**e**sti / št**ě**stí*
PAL	**palatalisation**	*Pra**hě** / Pra**ze***
PROT	**protetic v-**	***v**občas / občas*
DIA	**diacritics**	*d**á**rovat / d**a**rovat*
ASIM	**cons. assimilation**	*v**s**pomínat / v**z**pomínat*
NASIM	***voiced/voiceless cons.***	***z**e / **s**e známým filozofem*
ALT	**vowel alternations**	*sl**a**čina / sl**e**čna*
CNTR	**digraphs**	*v Par**ž**iži / Pa**ř**íži*
SIB	**sibilants**	*specifi**s**ké / specifi**c**ké*
CHAR	**other (1 char)**	*z kuchyn**i** / kuchyn**ě***

Morphology

NAFF

incompatible morph. suffix *do nemocnic**y**/nemocnic**e***

VBX

incorr. compound verb forms *bude líb**í**/líb**it***

RFL

incorr. use of reflexive pronouns ***m**oje/**sv**oje*

Syntax

DEP valency / dependency

*z kuchyn**i**/kuchyn**ě***

AGR agreement

*česk**ého** / česk**é** vlády*

COP copula

*Ale **-/je** třeba , abychom poznavali mnoho nových slov .*

COMPL object (form)

*to**mu**/to**ho** si hodně vážím*

SUBJ subject

*za poslední rok **já/-** jsem se seznámil s hodně lidmi*

WO word order

*aby **někdo nás** / **nás někdo** čekal doma*

Lexicon/Use

ASP **aspect** ***pomohla/pomáhala***

MOD **modal verbs** ***musí/mohou***

NEG **negation** *Nemůžu si vůbec **ne**stěžovat / stěžovat*

COIN **new (coined) word** *do **pivarny/pivnice***

FGN **foreign word** ***baboška/babička***

USE **incorr. gender etc.** *tramvaj**em**/tramvaj**í***

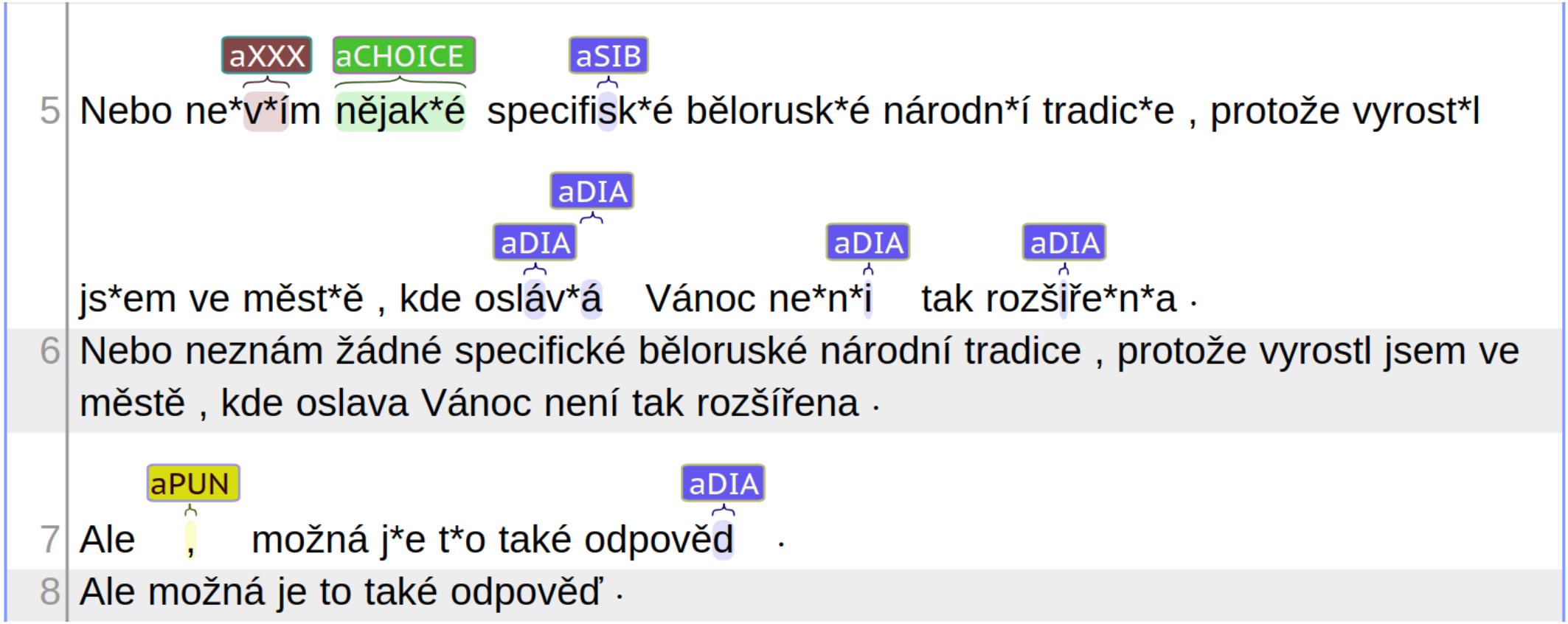
POS **incorr. POS** Ona je výborn**ě** / výborn**á** žena

PHR **incorr. use longer expressions**

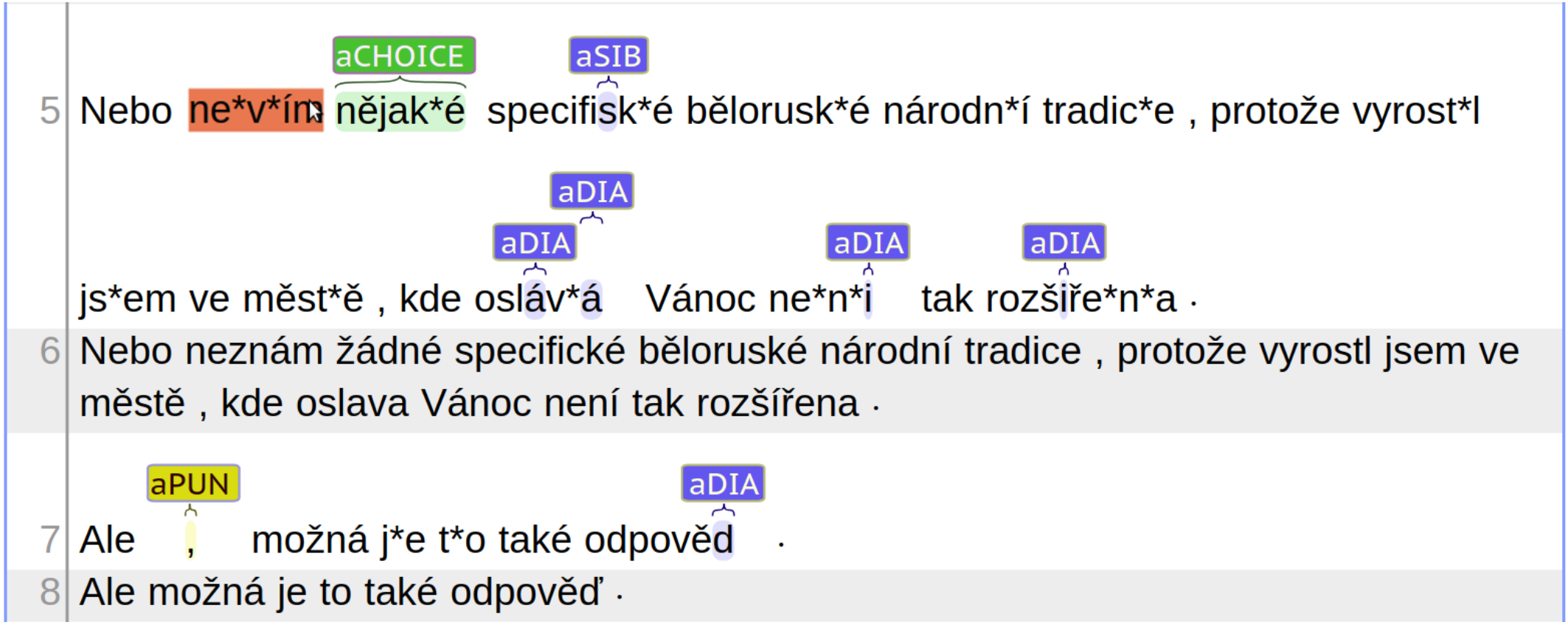
neměly rády vstávat** / **nerady vstávaly

CHOICE **incorr. choice of words** ***ví** / **zná***

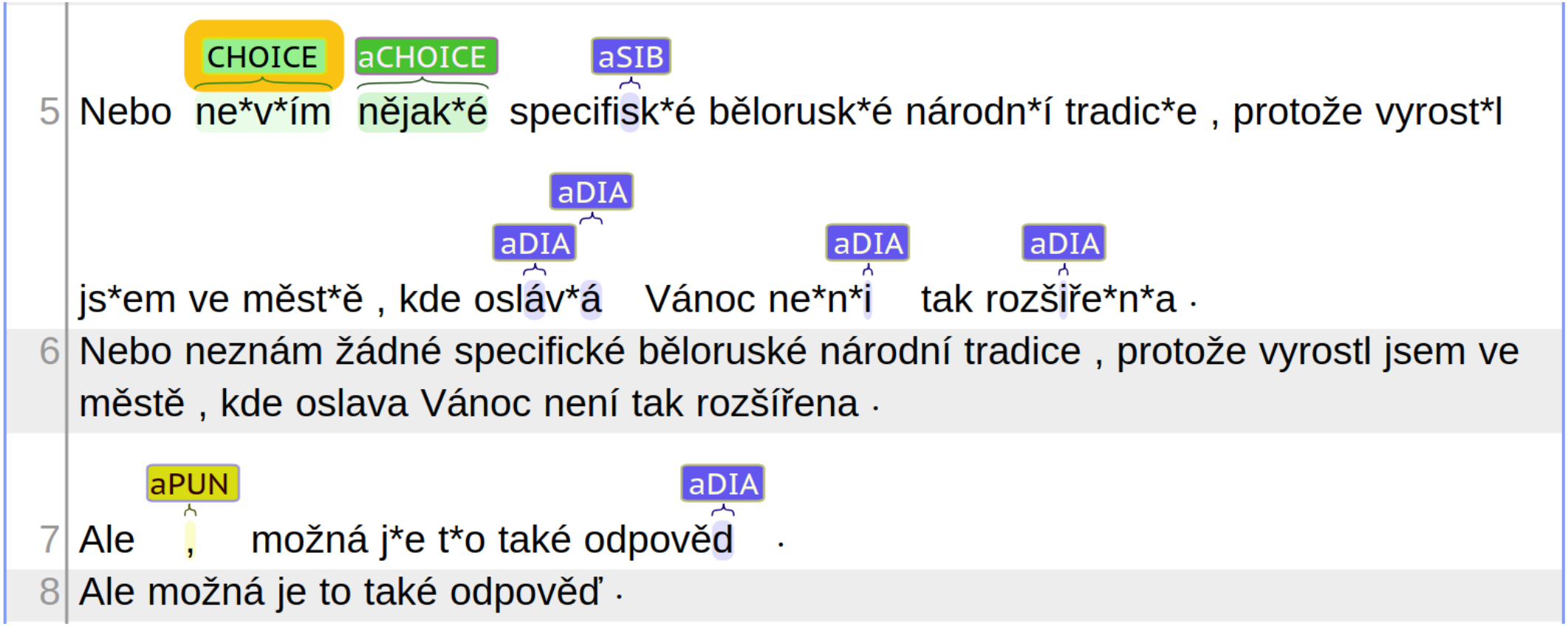
Brat annotation tool



Brat annotation tool



Brat annotation tool



Brat: data (file.txt)

Nebo ne*v*ím nějak*é specifisk*é bělorusk*é národn*í tradic*e , protože vyrost*l js*em ve měst*ě , kde osláv*á Vánoc ne*n*i tak rozšiře*n*a .
Nebo neznám žádné specifické běloruské národní tradice , protože vyrostl jsem ve městě , kde oslava Vánoc není tak rozšířena .
Ale , možná j*e t*o také odpověd .
Ale možná je to také odpověď .

Brat: data (file.ann)

T1 aXXX 163 166 v*í
T2 aCHOICE 168 175 nějak*é
T3 aSIB 183 184 s
T4 aDIA 261 262 á
T5 aDIA 264 265 á
T6 aDIA 277 278 i
T7 aDIA 287 288 i
T8 aPUN 428 429 ,
T9 aDIA 455 456 d

Text pre-processing

- **A simple morphematic analysis into prefix – root – suffix**

- **All differences between linked erroneous forms and emendations**

- **If possible, marked by one of 25 error tags (out of ORT, MPHON, SYN, LEX; to be checked by the annotator)**

Morphematic analysis

prefix – root – suffix

prefix and suffix: only those forming a new form inside the flex. paradigm e.g. ***po***jed*e* but ***přejed***e*

- **existing words:**

match possible suffixes for a given morph. tag with a list
e.g. *kamarádovi*: NNMS3 > *-i* and *-ovi* are matching,
choose *-ovi*

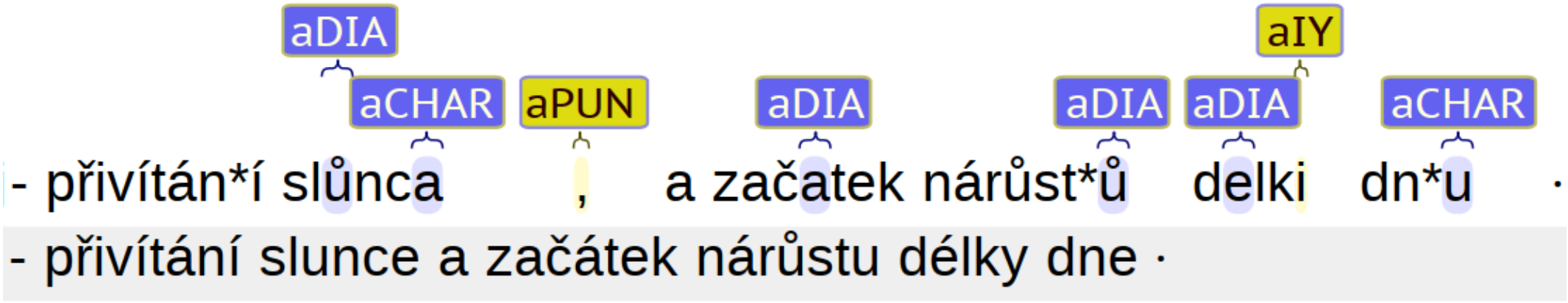
- **non-words:**

based on the corrected form (if similar enough)
e.g. *kamar**at**ov**y**/kamar**ád**ov**i***, suffix of *kamarádovi* is ***-ovi***, similar to ***-ovy***, choose ***-ovy***

Automatic error annotation

Find and mark differences between linked erroneous forms and emendations, assign error tags.

Create brat text and annotation file.



25 automatically assigned error tags (aDIA x DIA): to be checked by the annotators.

To do...

Manual annotation.

Extend and refine automatic annotation based on the results of manual annotation.

Experiment with using automatic text correction instead of manual annotation.

Perhaps one day, a fully automatic (simple) tool for CzSL text correction and error tagging.

References

Eckart de Castilho, R., Biemann, C., Gurevych, I. and Yimam, S.M. (2014): WebAnno: a flexible, web-based annotation tool for CLARIN. In Proceedings of the CLARIN Annual Conference (CAC) 2014, Soesterberg, Netherlands.

Jelínek, T., Štindlová, B., Rosen, A. and Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka et al. (eds), Text, Speech and Dialogue – Proceedings of TSD 2012, no. 7499 in LNCS, p. 127–134.

Richter M., Straňák P., Rosen A. (2012). Korektor – A System for Contextual Spell-checking and Diacritics Completion. In Kay M., Boitet C.: Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012). Mumbai, India, p. 1-12.

Rosen, A. (2016). Building and using corpora of non-native Czech. In B. Brejová (ed), Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016), vol. 1649 of CEUR Workshop Proceedings, Bratislava, Slovakia, p. 80–87.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou S. and Tsujii, J. (2012). Brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012, 102–107.

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