Syntactic annotation of a second-language learner corpus

Jirka Hana & Barbora Hladká

Charles University Prague

ICBLT 2018

CzeSL – Corpus of L2 Czech

CzeSL – Czech as a Second Language

- Part of AKCES – Acquisition Corpora of Czech
- Essays written by non-native speakers of Czech
- A1 – C1 CEFR proficiency levels

CzeSL – two releases

- CzeSL-SGT
  - 8,600 essays, 1.1M tokens
  - http://hdl.handle.net/11234/1-162, CC BY-SA-3.0

- CzeSL-man <– we work with this here
  - 645 essays, 120K tokens, 11K sentences
  - Manually corrected and annotated for errors
  - https://bitbucket.org/czesl, CC BY-SA-3.0

CzeSL – number of documents by CEFR Level

Level		**Documents**
Basic user	A1	57
	A1+	3
	A2	111
	A2+	145
Independent user	B1	176
	B2	124
Proficient user	C1	12
Unknown		17
Total		645

Non-native and native language are different

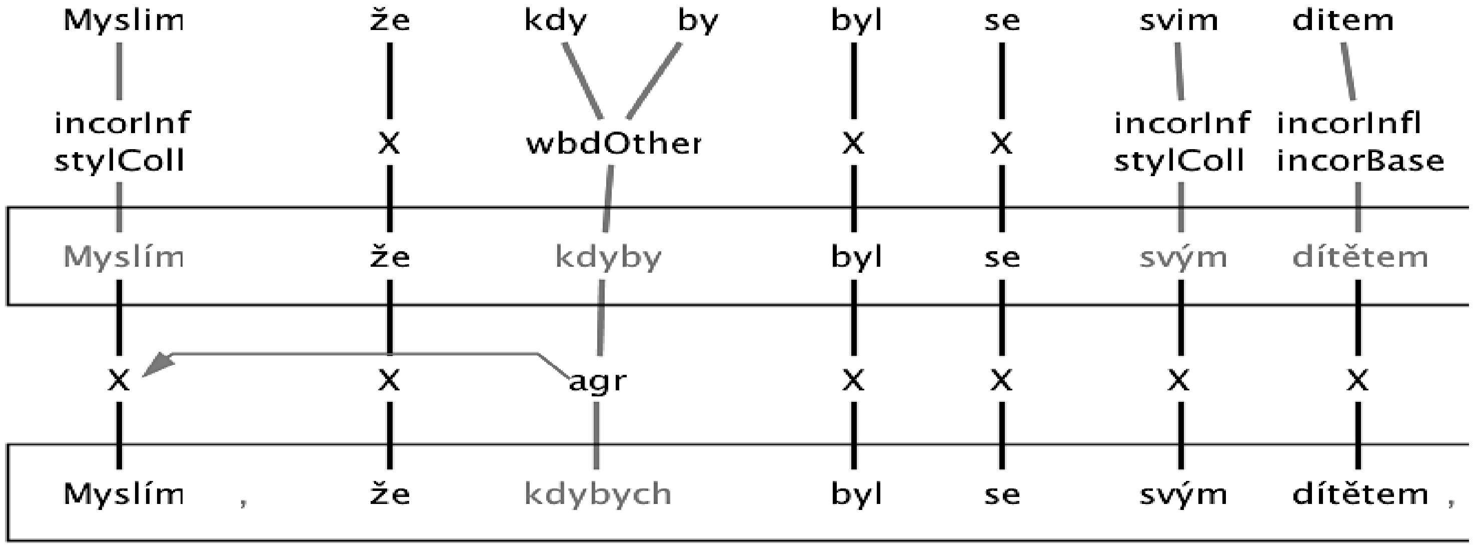
Non-native langue has:

- Errors in spelling, grammar, vocabulary, collocations
- Different distribution of vocabulary and syntactic constructions

CzeSL: Error Annotation Scheme

Tier 0	original text:	Myslim		že	kdy	by	byl	se	svim	ditem	...
Tier 1	words correct:	Myslím		že	kdyby		byl	se	svým	dítětem	...
Tier 2	contextually correct:	Myslím	,	že	kdybych		byl	se	svým	dítětem	...
		$\text{think}_{\text{SG1}}$		that	if_{SG1}		was_{MASC}	with	my	child	...
		\`I think that if I were with my child ....’									

corrections



Sample non-native text: My Family

**Jmenujese** Adam. **Ja** jsem Mongolska. **Mongolska ma** 21 **kraji**. Moje rodina je **hezka jeste velka**. **Mongolska je** 3000 **million lidi**. **Ma tradični píseňka,** taneční. **Mongolska tradicni píseňka** je **hezka**. **Ješte ma** ”Morin khuur”. Morin Khuur to je muzika. Ten **hezka tradični** pohádka, píseň. **Mongolska** má mnoho **tradiční svátík**. **Třiba** Naadam, Tsagaarsur. **Ješte** mnoho **Velbloud**, **Kůn**, **Kravá,** **Koza**, **Ovce**. **Mongolsky** lidi dobrý. Mongolsko **ma** mnoho **hory** a **nemam ocean**. **Mongolska** hlavní **naměsto**. Ulaanbaatar.

ADAM, 18 Let

**Bydlim** v **Cechagh** už 6 **měsíc**.

Task: Annotate some structure of L2 Czech

Motivation:

- better understanding of L2 Czech (including its grammar)
- better computational processing of L2 Czech

Some structure?

- the deeper, the better, ideally semantics
- dependency syntax for practical purposes

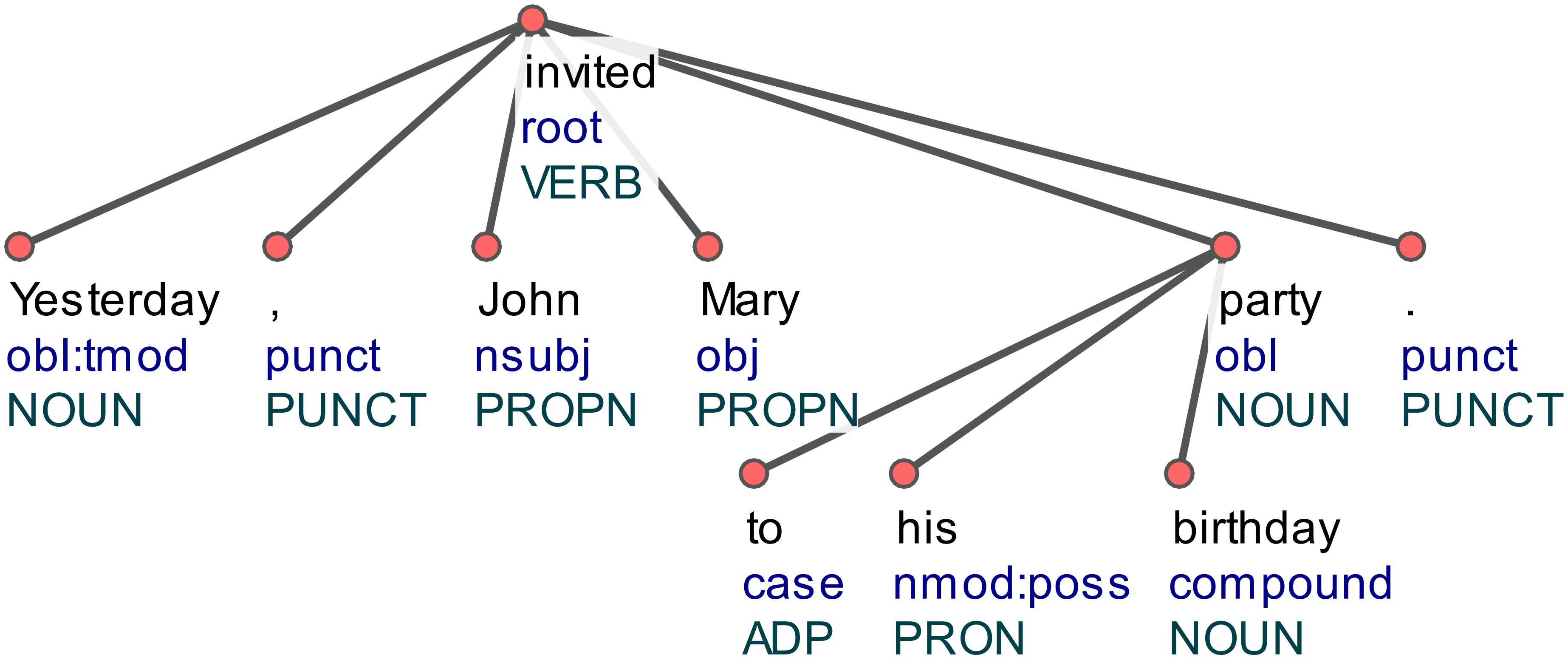
Work in progress ...

Universal Dependencies

Universal Dependencies (UD)

- 100+ corpora in 60+ languages
- syntactic annotation based on dependency syntax
- language agnostic (mostly)

Universal Dependencies (UD)



Yesterday, John invited Mary to his birthday party.

UD Annotation of Non-native Czech

Annotating original text

- Annotate the original text, not corrections
- Ideal case: use grammar of author‘s interlanguage
- Reality: often, not enough data
- Be conservative, assume as little as possible

Annotating original text – Example 1

- Standard – oblique (adjunct):

Vstoupit	do	místnosti.
enter	into	room.

`Enter a room.’

- Non-native – direct object:

Vstoupit	místnost.
enter	room.

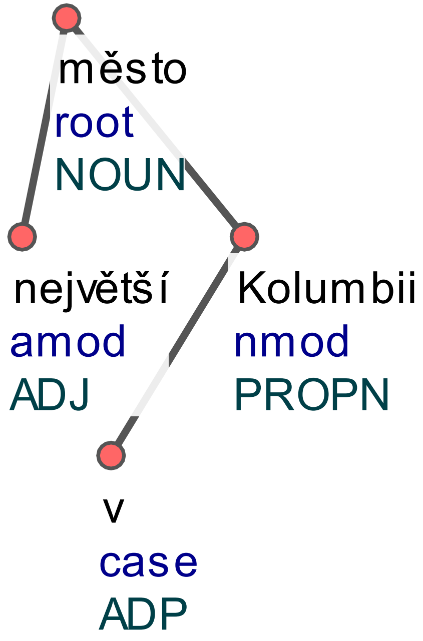
Intended: `Enter a room.’

Annotating original text – Example 2

- Standard – *nej* `most‘ is a prefix

největší	město	v	Kolumbii
biggest	city	in	Columbia

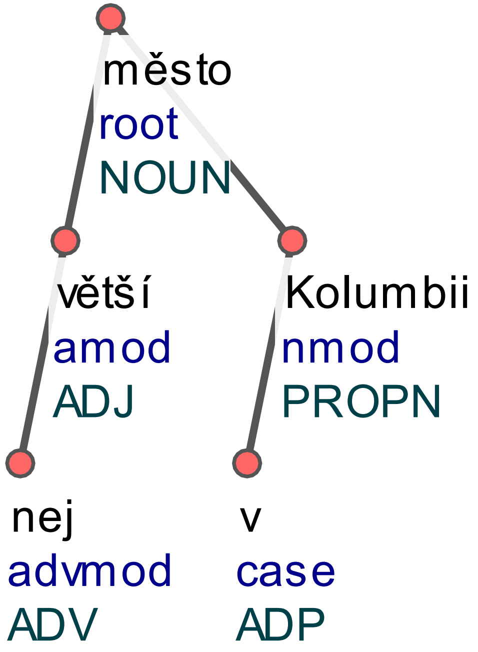
`the biggest city in Columbia’


město
root
NOUN
největší
amod
ADJ
Kolumbii
nmod
PROPN
v
case
ADP

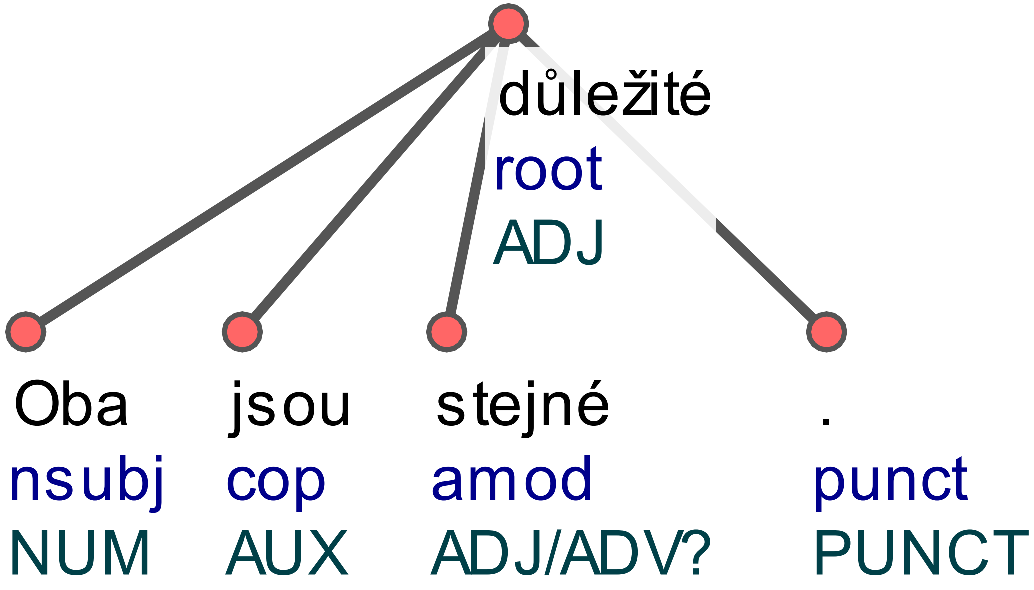
- Non-native – *nej* is a word:

nej	větší	město	v	Kolumbii
most	bigger	city	in	Columbia

`the biggest city in Columbia’


město
root
NOUN
větší
amod
ADJ
Kolumbii
nmod
PROPN
nej
advmod
ADV
v
case
ADP

Annotating original text – an unclear example



Oba	jsou	stejné	důležité.
both	are	equal	important

`Both are equally important’

- *stejné* = equal (adjective)
- *stejně* = equally (adverb)

What is it?

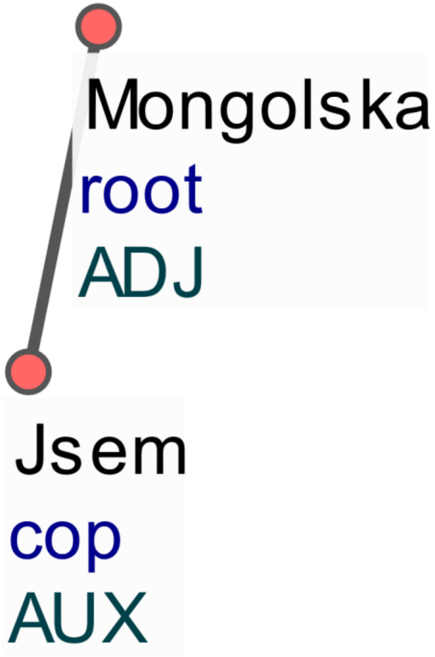
- spelling error
- adj/adv neutralization
- other error

What is the lemma, POS, syntactic fnc?

Sometimes UD helps …

Jsem Mongolska.

`I am Mongolian / a Mongolian / from Mongolia’



- *Jsem mongolský.* – adjective, not in std language
- *Jsem Mongol.* – inhabitant, noun
- *Jsem z Mongolska.* – country, preposition + noun

The same structure in UD

Current status

- 2,100 sentences out of 11,000 annotated so far
- 100 sentences double annotated with Cohen’s kappa:
  - Universal POS: 0.93
  - Dependency Label: 0.89
  - Relation: 0.93

Future work

- More double annotated data
- More annotated data – annotate the whole CzeSL
- Test standard and custom trained parsers