Universal Dependencies and non-native Czech

Jirka Hana & Barbora Hladká

Charles University Prague

TLT 2018

CzeSL – Czech as a Second Language

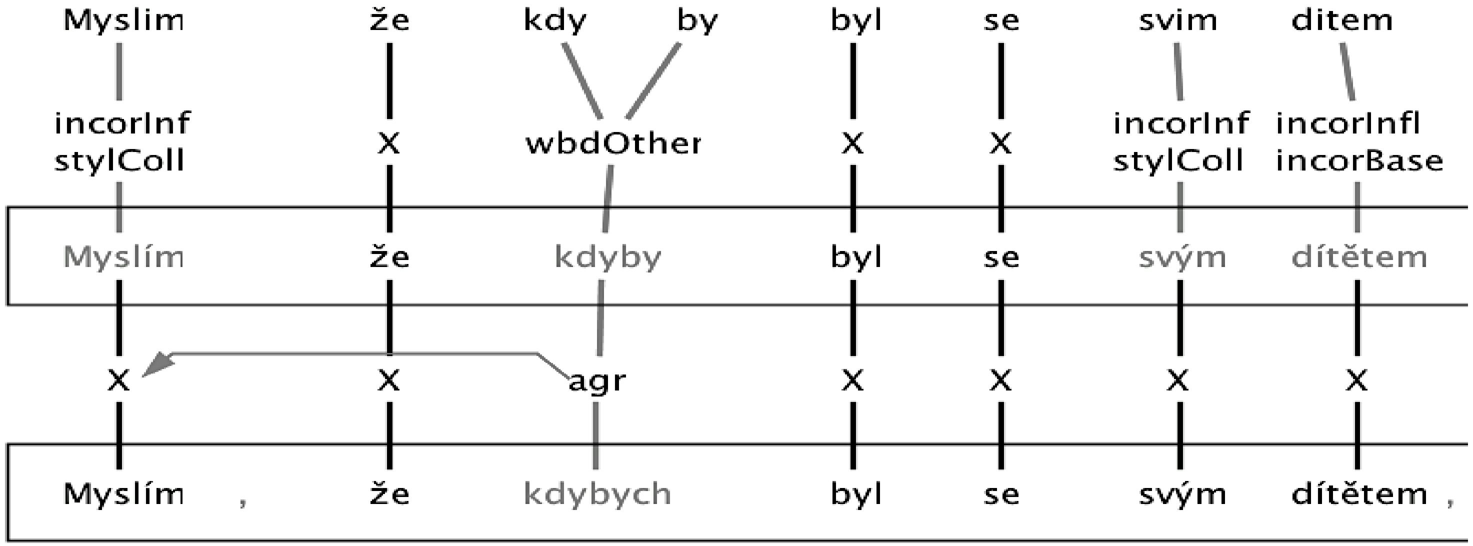
- Texts written by non-native speakers of Czech

- CzeSL-man subcorpus <– we work with this here
    - 645 essays, 120K tokens, 11K sentences
    - A1-C1 CEFR proficiency levels
    - Manually corrected and annotated for errors
    - https://bitbucket.org/czesl, CC BY-SA-3.0

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	...
		think$_{SG1}$		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 [Name]. **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.

[NAME], 18 Let

Bydlim v **Cechagh** už 6 **měsíc**.

Task: Annotate some structure of L2 Czech

Motivation:

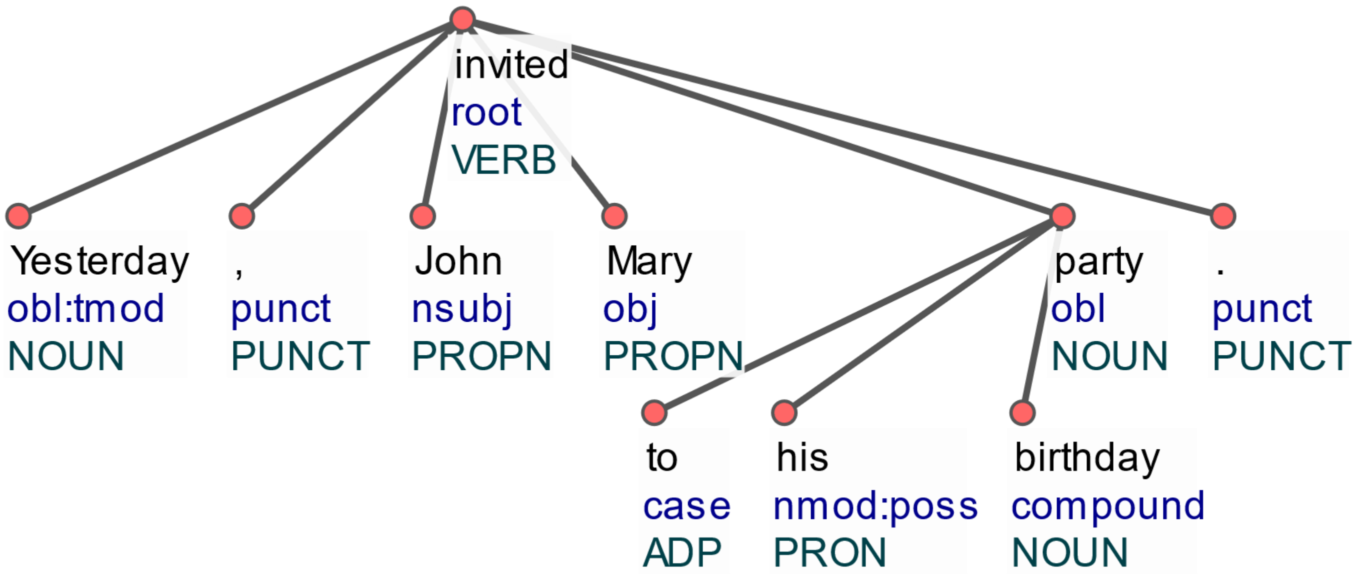
- better understanding of L2 Czech (including its grammar)
- better computational processing of L2 Czech

Some structure?

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

Universal Dependencies (UD)

- dependency-based syntactic annotation
- language agnostic (mostly)
- v. 2.3 (Nov 2018)
  - 129 treebanks
  - 76 languages



Yesterday, John invited Mary to his birthday party.

Approach

- Annotate the original text, not corrections
- Ideal case: use grammar of author‘s interlanguage
- Reality: often, not enough data

Trees – Example 1: oblique vs direct object

- Standard – `obl`

Vstoupit	do	místnosti.
enter	into	room.

`Enter a room.’

- Non-native – `obj`

Vstoupit	místnost.
enter	room.

Intended: `Enter a room.’

Trees – Example 2: superlative

- Standard – *nej* `most‘ is a prefix

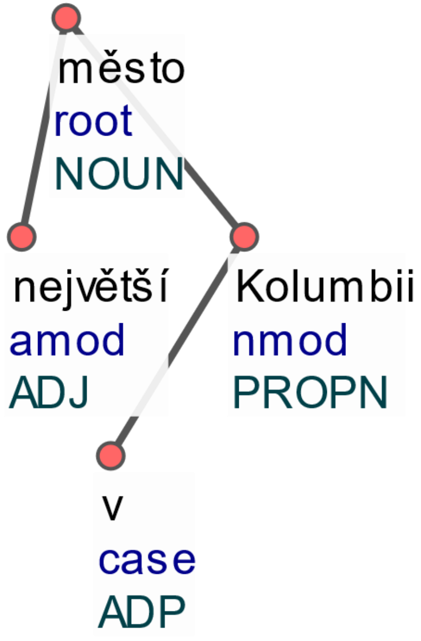
největší	město	v	Kolumbii
biggest	city	in	Columbia

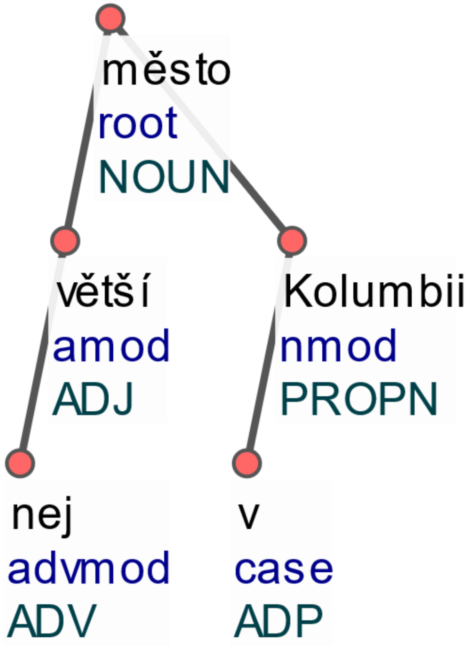
`the biggest city in Columbia’

- Non-native – *nej* is a word

nej	větší	město	v	Kolumbii
most	bigger	city	in	Columbia

`the biggest city in Columbia’





Trees – Example 3: quantifiers

- Standard

	quantifier + genitive pl	agree in case
DET	`det:numgov` *mnoho tygrů* ‘many tigers’	`det:nummod` *s mnoha tygry* ‘with many tigers’

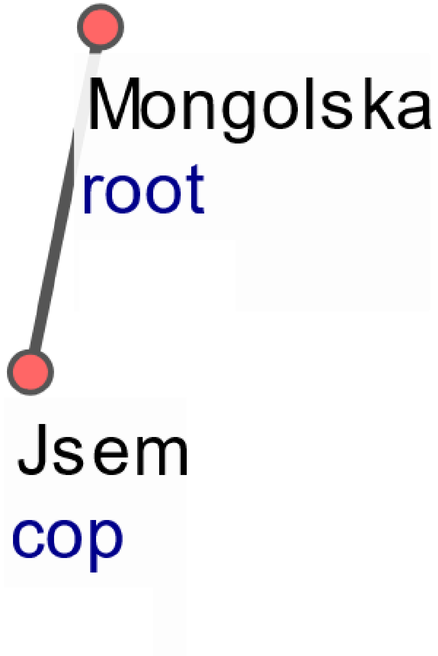
- non-native – quantifier + nominative:
  - *Mongolska má mnoho **tradiční svátík**.*
    `Mongolia has many traditional holiday’
  - *Ješte mnoho **Velbloud**, **Kůn**, …*
    `Also many Camel, Horse, ...’

⇒ Simpler situation: `det:nummod` / `nummod` (similar to English)

Sometimes UD helps

Jsem Mongolska.

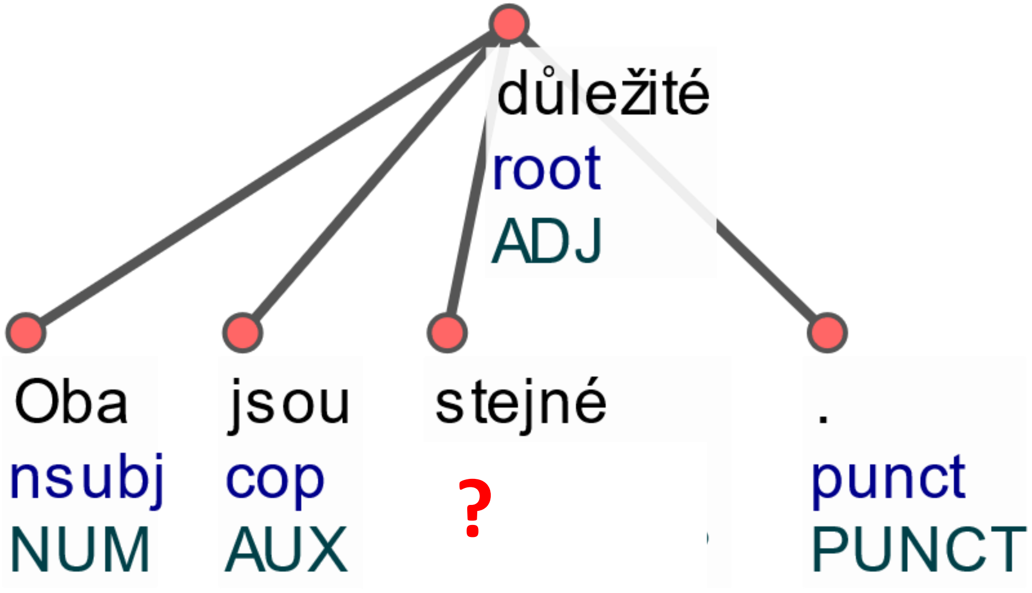
`I am Mongolian / a Mongolian / from Mongolia’



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

The same structure in UD

Sometimes not: ADJ/ADV neutralization



Oba	jsou	stejné	důležité.
both	are	equal	important

`Both are equally important’

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

	ADJ	ADV
amod	standard UD	?
advmod	?	standard UD

Ideally:

- POS = ADJ/ADV
- dependency = *MOD

But need sufficient evidence, can be:

- spelling error
- morphology

Sample non-native text: My Family

Jmenujese [Name]. **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.

[NAME], 18 Let

Bydlim v **Cechagh** už 6 **měsíc**.

Non-native sentences

	grammatically correct	grammatically incorrect
easy to understand	A	B
hard to understand	C	D

Conclusion

- Be conservative, assume as little as possible
- UD sometimes forces us to make unwarranted decisions

- Semantic annotation might be the right thing

Current status

- 2,200 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