Parsing Texts of Non-Native Czech

Jirka Hana & Barbora Hladká

Charles University Prague
NLPTEA 2017

CzeSL – Czech as a Second Language

- Part of AKCES – Acquisition Corpora of Czech
- Essays written by non-native speakers of Czech
- A1 – C2 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
  - Manually corrected and annotated for errors
  - https://bitbucket.org/jhana/czesl, CC BY-SA-3.0

CzeSL – number of texts 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

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

Work in progress

Ultimate: Create high-level NLP tools for non-native Czech

Intermediate: Create a parser for non-native Czech

For now: Identify top level syntactic functions
(Predicate, Subject, Object, Nominal predicate)

What is the problem

- Syntactic annotation is expensive
- For non-native language, it is not always clear what the gold standard should be.
- Can non-native parser be trained on native text?

What did we do

- Annotated a corpus of non-native Czech (CzeSL) with selected syntactic functions
- Trained an MST parser on
  - a corpus of standard Czech (PDT)
  - a subset of standard Czech corpus with simple sentences (STYX)
- Evaluated the results

Syntactic annotation of CzeSL-man

- Only main syntactic functions are distinguished
  - Predicate
  - Subject
  - Object
  - Nominal predicate

Syntactic annotation of CzeSL-man

- Annotation of original text not of target hypothesis
- Standard:

Vstoupit	do	místnosti.
enter	into	room.

`Enter a room.’

Annotated as
directional adjunct (obl in UD)

- Possible non-native:

Vstoupit	místnost.
enter	room.

Intended: `Enter a room.’

Annotated as
object (obj in UD)

But some interpretation is needed …

Jsem Mongolska. `I am Mongolian / from Mongolia'

Could be viewed as:

- *Jsem mongolský.* – adjective
  - nominal predicate (UD: root)
  - not in standard language
- *Jsem Mongol.* – inhabitant, noun
  - nominal predicate (UD: root)
- *Jsem z Mongolska.* – country, preposition + noun
  - adjunct (UD: obl)

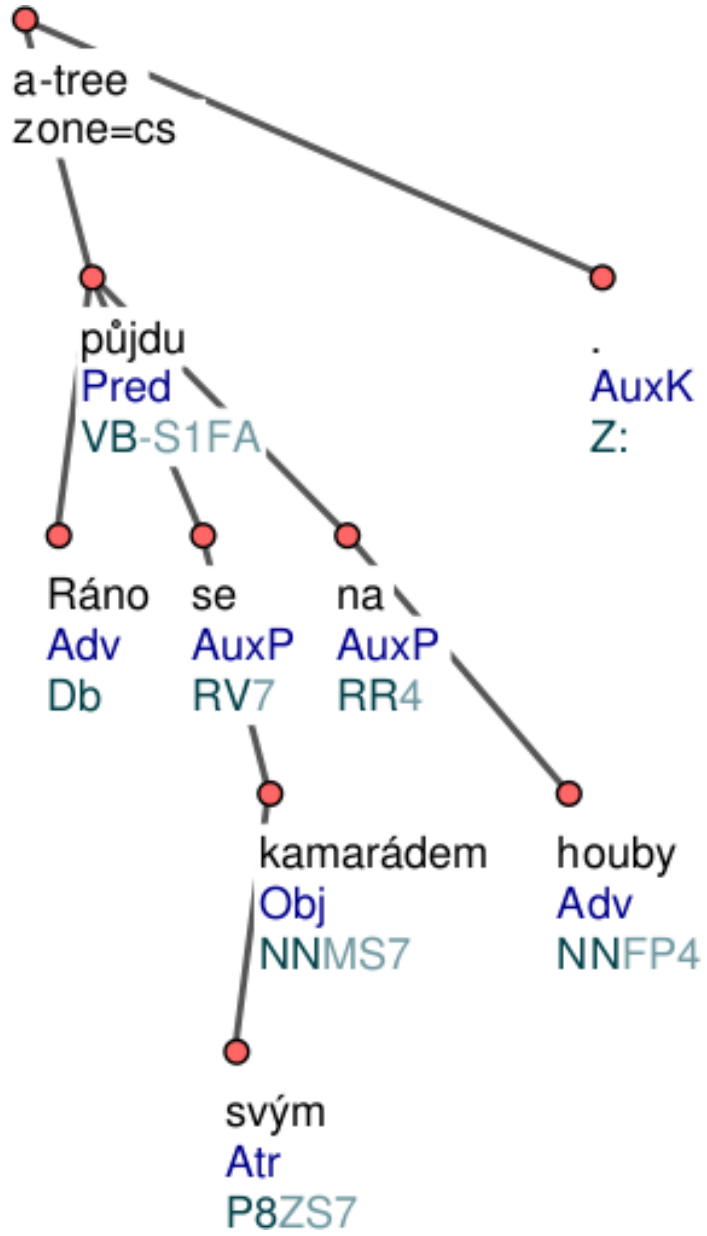
Prague Dependency Treebank (PDT)

- Czech Newspaper texts
- Annotated for morphology, dependency, ...
- Example:

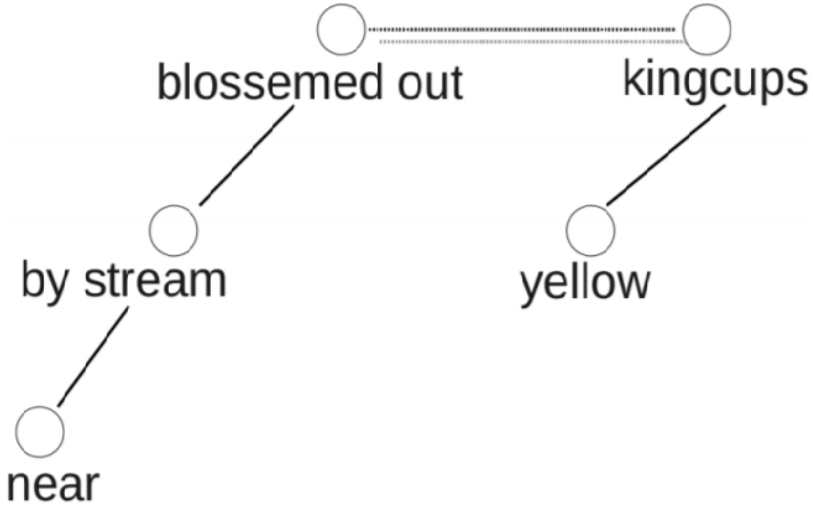
Ráno	půjdu	se	svým	kamarádem	na houby	.
in-the-morning	I-will-go	with	my	friend	mushrooming	.

`I will go mushrooming with my friend in the morning.'

- Available: http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3

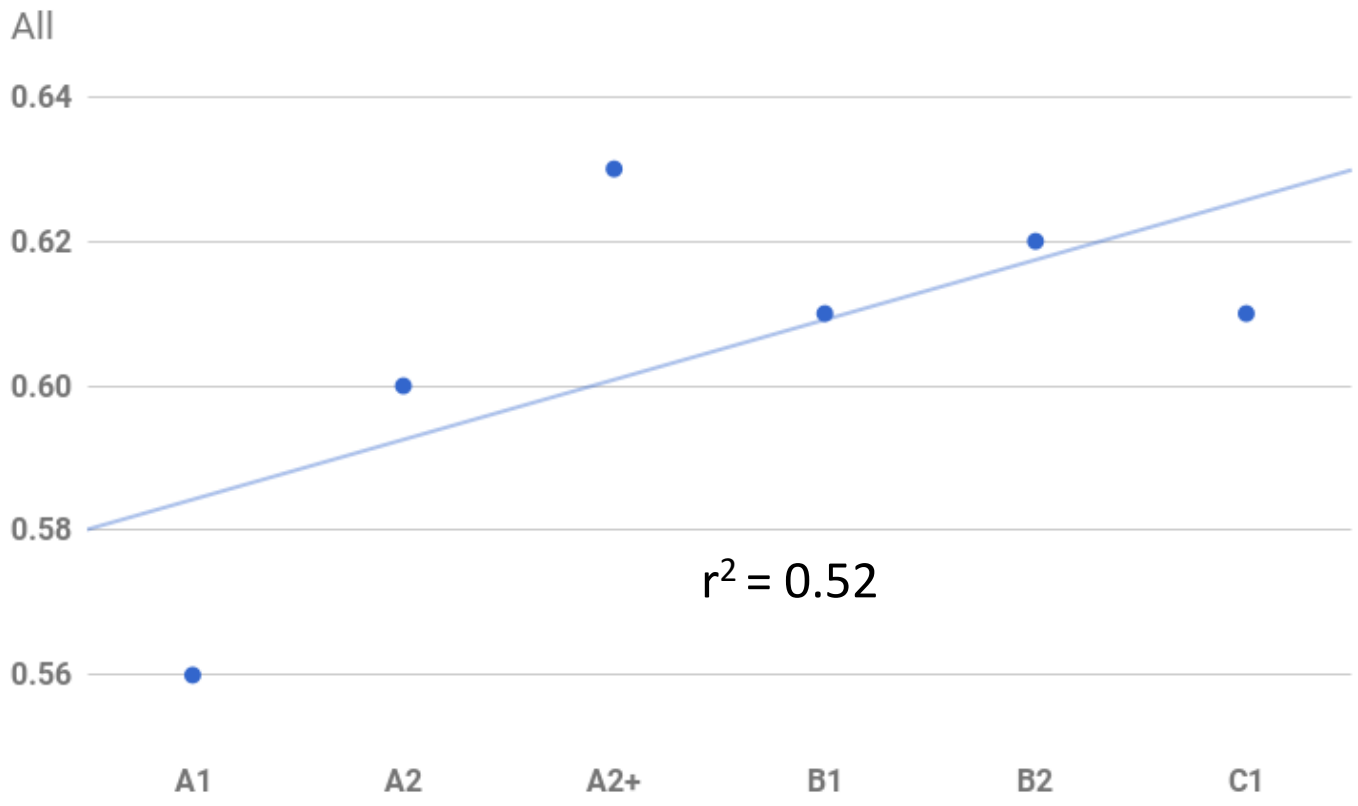


STYX

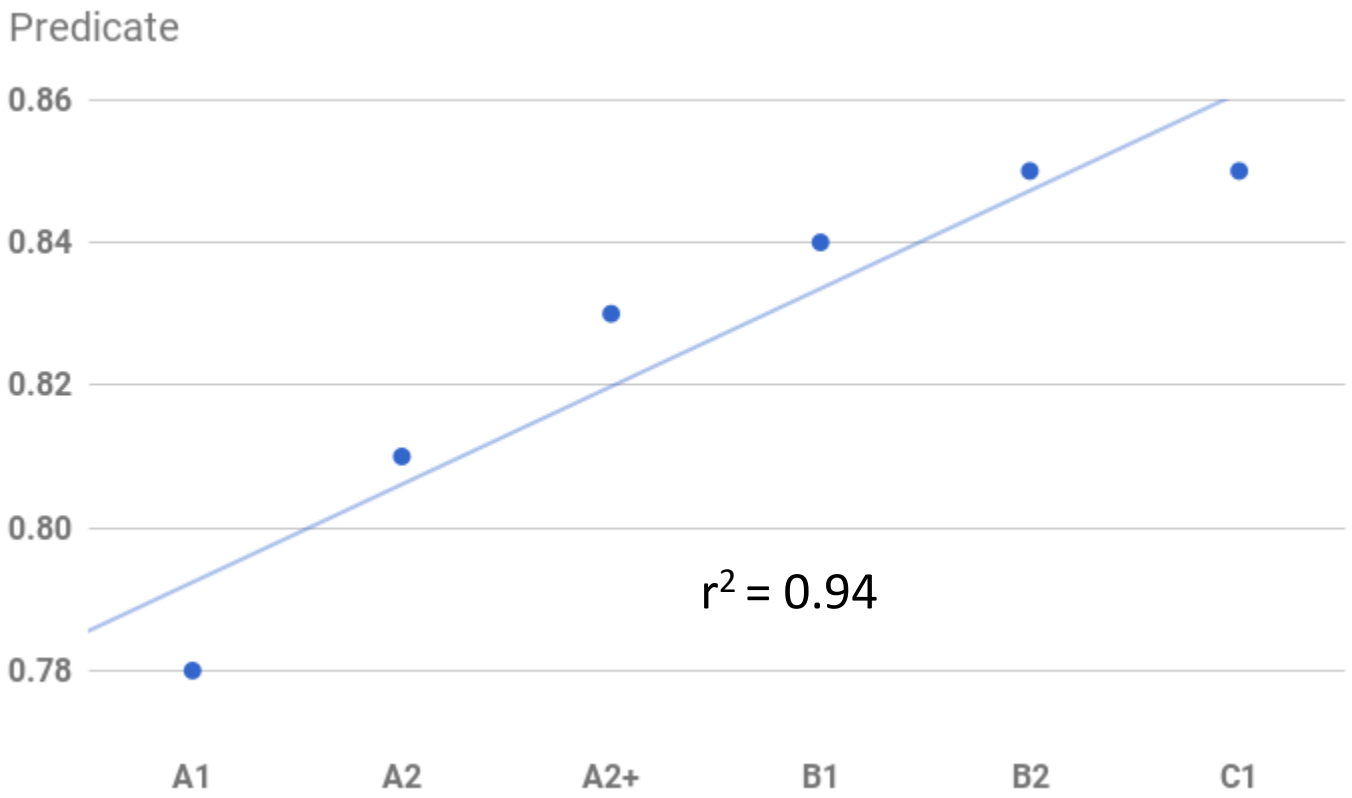


- Simpler language subset of PDT
- Complex phenomena discarded
- Rule-based transformation of PDT annotation into Czech school annotation
- Available: http://hdl.handle.net/11234/1-2391

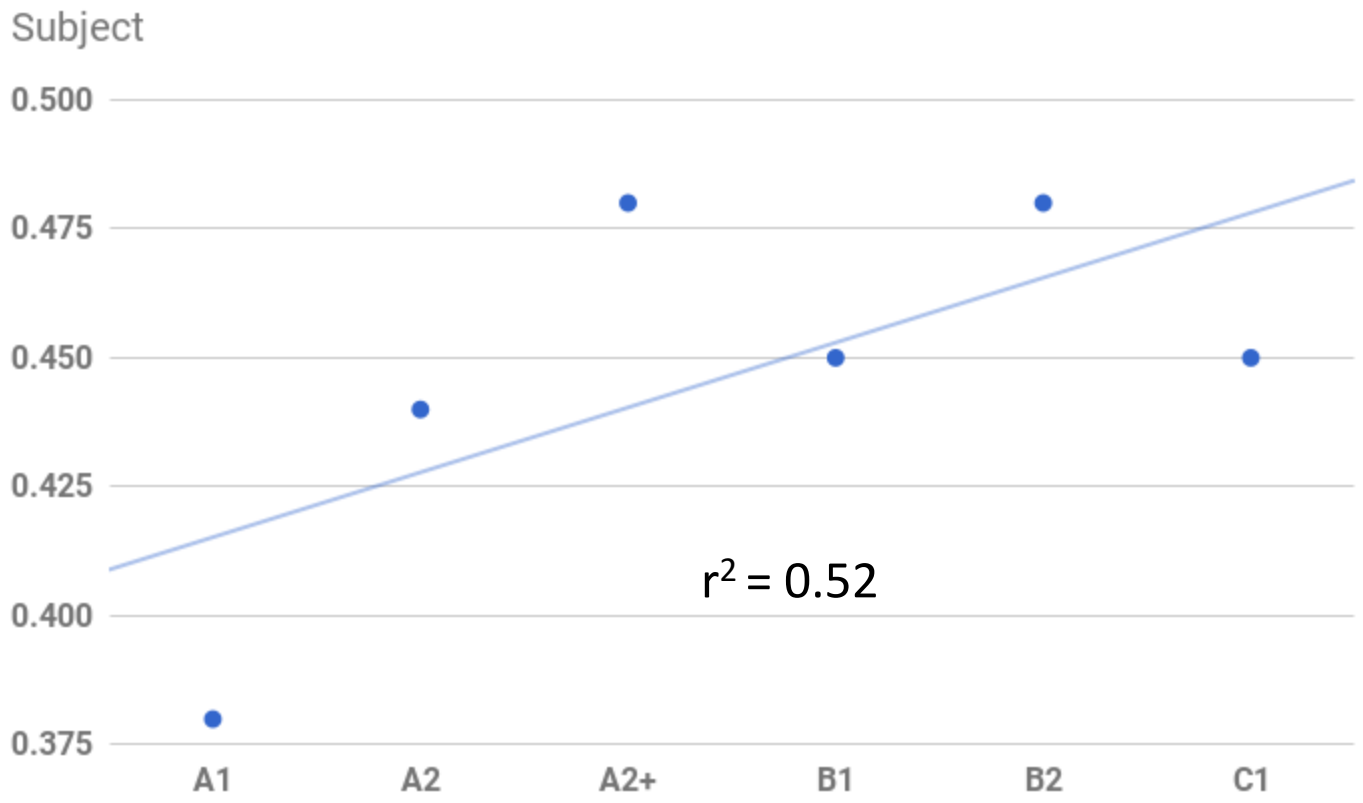
Results – Parser trained on PDT – All



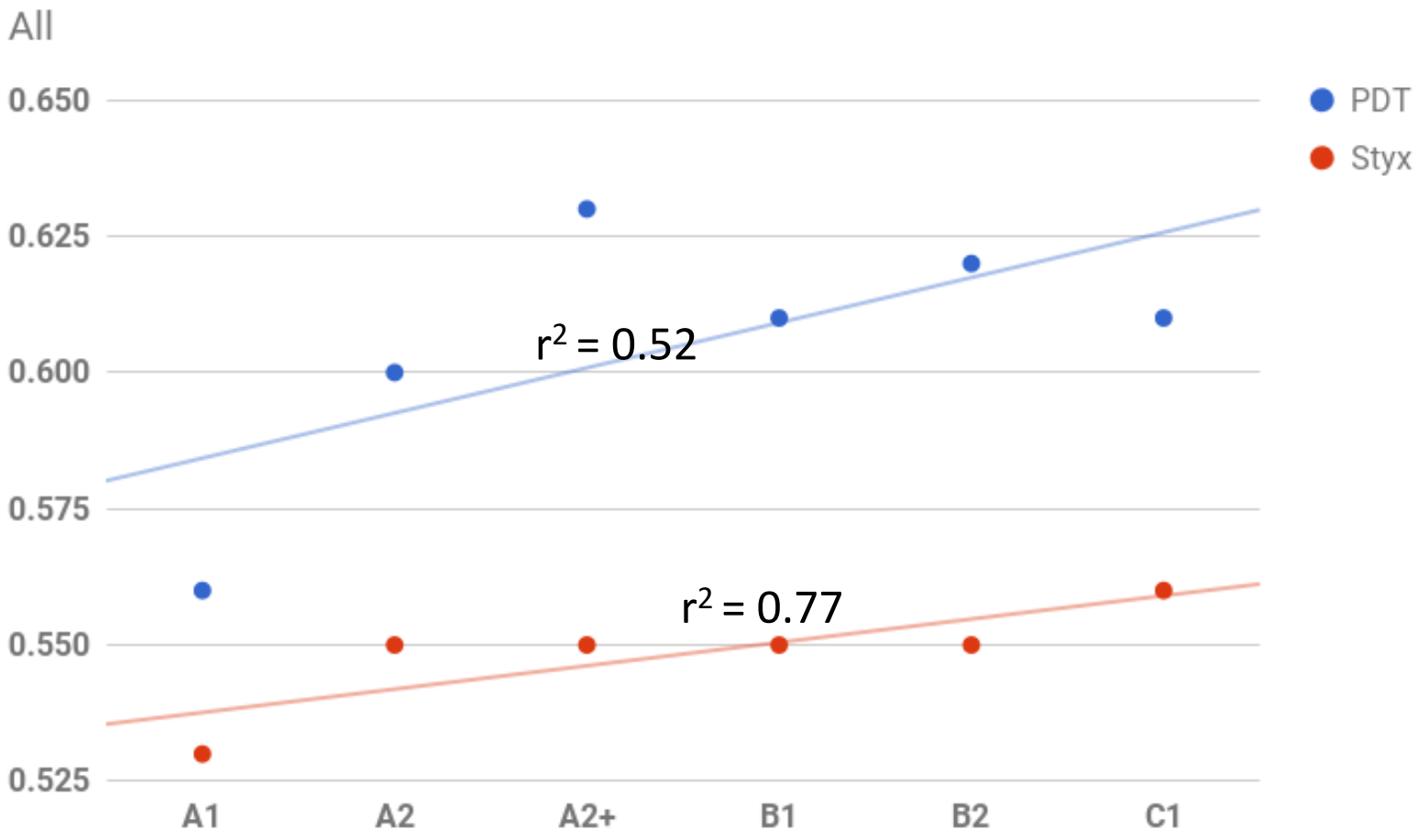
Results – Parser trained on PDT – Predicate



Results – Parser trained on PDT – Subject



Results – Parsers trained on PDT/Styx – All



Conclusion

- parser trained on native language corpus works for non-native corpus (at least on basic function labels)
- using simpler native language does not seem to help

Future work

- (nearly) full UD annotation
- two parallel annotations