Can Surprisal Predict Mutual Intelligibility of Slavic Languages?

Klára Jágrová, Tania Avgustinova, Irina Stenger, Andrea Fischer

{kjagrova, avgustinova}@coli.uni-saarland.de
{ira.stenger}@mx.uni-saarland.de
{afischer}@lsv.uni-saarland.de

Context: a) Receptive Multilingualism

- **Slavic languages** – expected to be mutually intelligible
- **Reading intercomprehension:** depending on language combination and various factors: orthography, lexis, morphology, syntax
- semantic and/or syntactic context: helpful, neutral, or impairing

b) Language Modeling

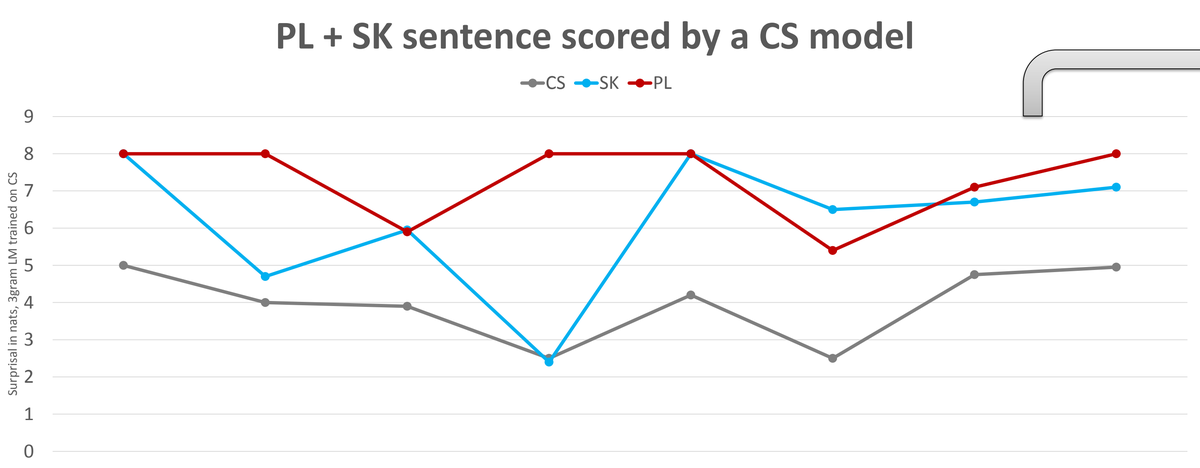
- **Surprisal** informs about the (un-)predictability of certain words in monolingual sentential context

$$Surprisal(unit) = \log_2 \frac{1}{P(unit \mid \text{Context})}$$

- measures information density
- correlates with cognitive effort

Aim

- **Modeling a Slavic reader** who is trying to decode a related but unknown foreign language
- compare predictions of models to actual human performance (experiments to be launched)



PL: W Parlamencie Europejskim wszystkie języki urzędowe są równie istotne.
SK: Všetky úradné jazyky sú v Európskom parlamente rovnako dôležité.
CS: V Evropském parlamentu mají všechny úřední jazyky stejný význam.

- with this method: linguistic distance displayed only as a tendency

- divergent orthography
- divergent morpho-syntax } OOV problem → loss of predictive power
- divergent lexis
- divergent word order → interesting!

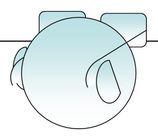
→ solution of OOV problem: analyse the factors involved separately
→ Which difficulties in reading intercomprehension are caused by divergent word order?

Overall aim:
→ Which units of language carry which information in concrete intercomprehension situations?

1) Evaluating readers' responses from free translation experiments

TRANSPARENCY ↓ / SURPRISAL →

	HR stimulus:	Daleko	je	kuća	moja.
readers	BG:	Daleko	e	kăŝa	moja.
	CS:	Daleko	je	chata / dům	moje / můj.
	PL:	Daleko	jest	chata / dom	moja / moj.
	RU:	Daleko		xata / dom	moja / moj.
		[Far away	is	sth. ♀	of mine.]



Influencing readers' choice of translation:

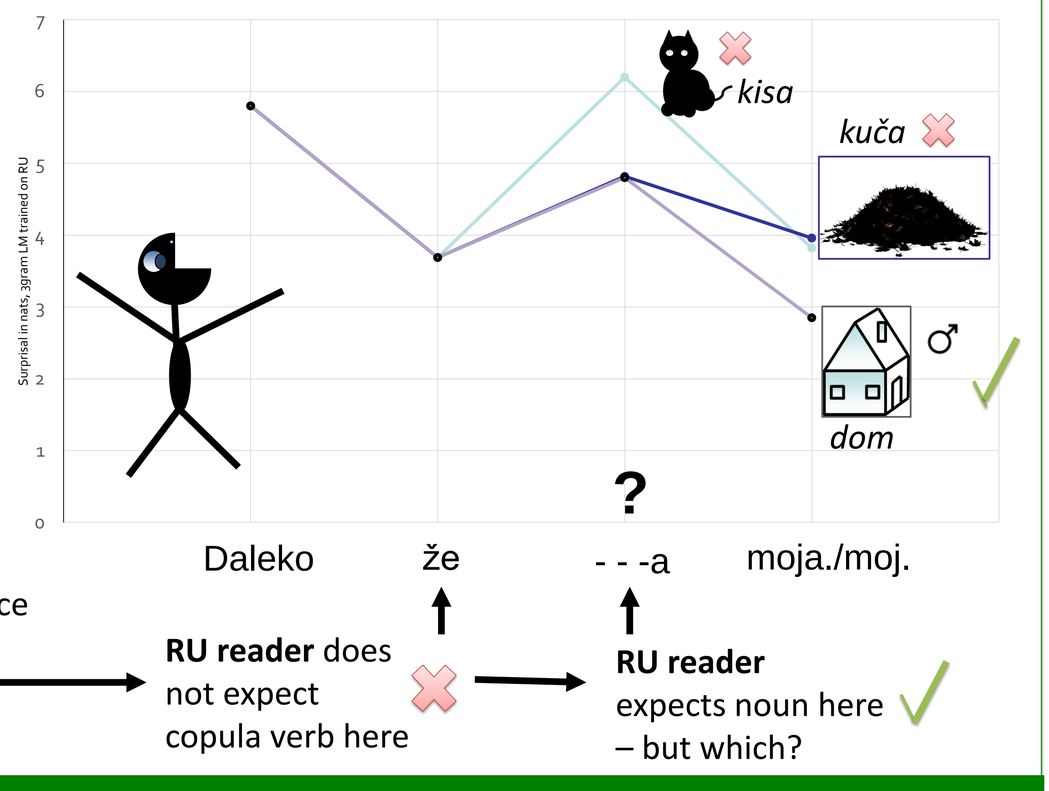
- initial letter
- letter similarity to L1
- orthographic distance to words in L1
- phonetic similarity to L1 (assumed by reader)
- lexical interference from L1
- grammatical gender
- neighbourhood density of words
- **predictability of the word in L1 context**
- …

Responses by RU readers:

a) *Далеко **же киса** моя.*
*/daleko **že kisa** moja/*
script not transparent → interference

b) *Далеко же **куча** моя.*
*/daleko že **kuča** moja/*
phonetic similarity → lexical interference

Comparison: surprisal of answers by RU readers



RU reader does not expect copula verb here → **RU reader** expects noun here – but which?

2) Predicting processing effort for readers in free translation experiments

Example: PL sentence read by a CS reader

PL: ***W Parlamencie Europejskim wszystkie języki urzędowe są równie istotne.***

[In Parliament European all languages official are equally important.]

1) Removing differences in orthography + morphology + lexis towards CS:

CS: ***V Parlamentu evropském všechny jazyky úřední jsou stejně významné.***

premodifier vs. postmodifier; position of clitics → **models**

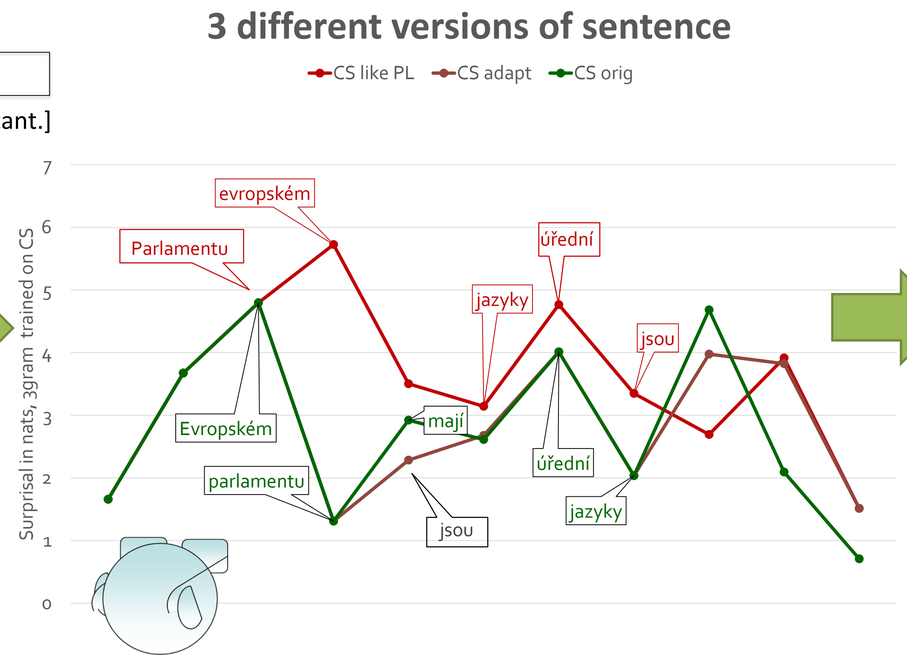
2) Rearranging sentence in order to obtain syntactically correct CS:

V Evropském parlamentu jsou všechny úřední jazyky stejně významné.

copula vs. light verb; ADVP vs. NP

3) Comparison: CS original translation (EuroParl corpus)**:**

V Evropském parlamentu mají všechny úřední jazyky stejný význam.



- Slavic: linearisation in clausal domain is free – depends mainly on topic focus
- Higher suprisal → higher processing effort expected for morphosyntactic phenomena that pose grammatical constraints, e.g. here:
  - clitics in Wackernagel position
  - premodification vs. postmodification
- In progress: representative collection of such morphosyntactic phenomena

Source of example sentence:
http://www.europarl.europa.eu/aboutparliament/en/20150201PVL00013/Multilingualism/ [retrieved: 28.08.2016]