C4: Mutual Intelligibility and Surprisal in Slavic Intercomprehension

Andrea Fischer, Klára Jágrová, Irina Stenger
Tania Avgustinova, Dietrich Klakow, Roland Marti

Slavic languages: intercomprehensible to various degrees

Objective: find mechanisms of linguistic coding + statistical evidence of mutual intelligibility

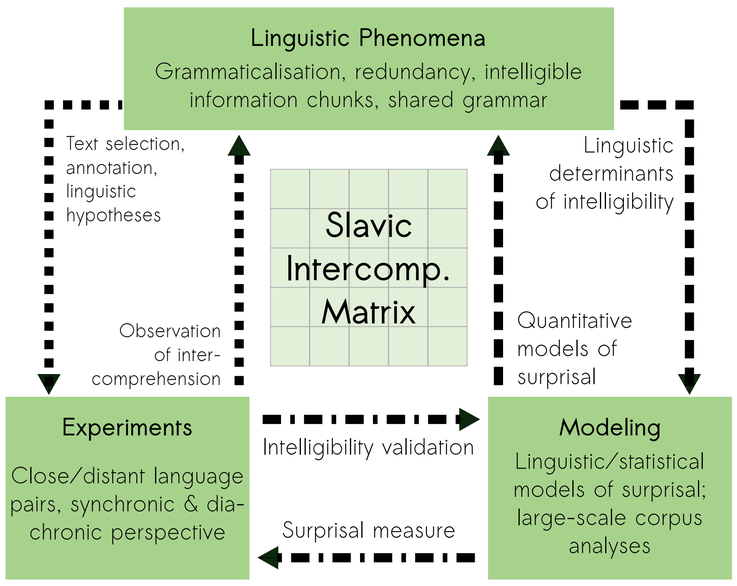
Source: http://en.wikipedia.org/wiki/Slavic_languages

Focus: reading intercomprehension

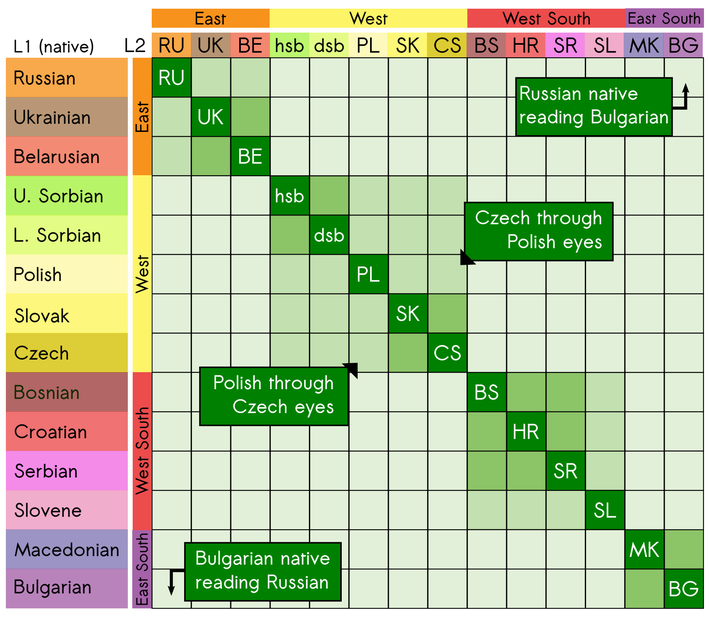
Aspects: orthography, morphology, lexis, syntax, semantics

Methods: statistics, language modeling, machine translation information theory, Slavic linguistics

Research Plan



Slavic Intercomprehension Matrix



Receptive Multilingualism

Special mode of language use: understanding unknown languages without being able to use them for speaking or writing. How and why is this possible?

(PL) W Parlamencie Europejskim wszystkie języki urzędowe są równie istotne.
(CS) V Evropském parlamentu mají všechny úřední jazyky stejný význam.
(BG) В Европейския парламент всички официални езици са еднакво важни.
(RU) В Европейском парламенте все официальные языки в равной степени важны.
(EN) In the European Parliament, all official languages are equally important.

Example sentence taken from http://www.europarl.europa.eu/aboutparliament

Meaningful Units of Language

Certain **constructions** encode specific information

V Evropském *parlamentu* ... (CS), meaning: "parliament"

Noun, singular, male, locative case, preceded by adjective, part of PP

(RU) В Европейском парламенте — -ом + -е: prepositive case
(BG) В Европейския парламент — -ия: determiner (male adjective short)

Identifying Encoding Schemes of Natural Languages

Objective: well-founded **statistical model** of natural language understanding
→ fundamental advance in computational linguistics research

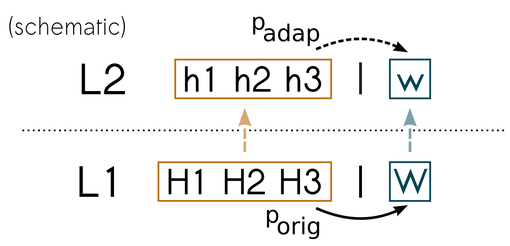
We expect:
- Diminished intelligibility through missing units
- Confusion through mis-recognition of units

→ discover informative elements of natural language

Modeling: Language as Domain

Basic idea: surprisal of statistical **n-gram language models** correlates with cognitive effort, but n-grams need to be adapted to process a different language

Smith, Nathaniel J., and Roger Levy. 2013.
The Effect of Word Predictability on Reading Time Is Logarithmic.
In Cognition 128.3 (2013). 302-319.



Decoding as Domain Adaptation

Explicit "latent" space describing each language
Decompose words into meaningful units
→ decode the words from unknown languages by similarity to known units
→ treat them exactly as in-language words would be

Soft Class Language Model for Adaptation

N-gram class language model
→ relax notion of hard classes to soft ones → *features*

$$p(w \mid h) = \underbrace{\sum_{f \in F(w)} p(w \mid f)}_{\text{word is mixture of features}} \overbrace{\sum_{f_h \in F_X(w)} p(f \mid f_h)}^{\text{feature importance in context}} \underbrace{\prod_{i=1}^{N} p(f_{h_i} \mid w_{h_i})}_{\text{feature importance in word}}$$

Each individual word is agglomerate of meaningful units: list of features
→ each feature contributes individually to the word's identity

Preliminary Results: Orthography

Diachronically-based assumptions tested on parallel list of Pan-Slavic vocabulary for each language pair (high cognate rate)

English	Czech	Polish	Bulgarian	Russian
'horse'	kůň	koń	кон	конь
'body'	tělo	ciało	тяло	тело
'sea'	moře	morze	море	море
'brush'	štětka	szczotka	четка	щётка
'head'	hlava	głowa	глава	голова
'cow'	kráva	krowa	крава	корова

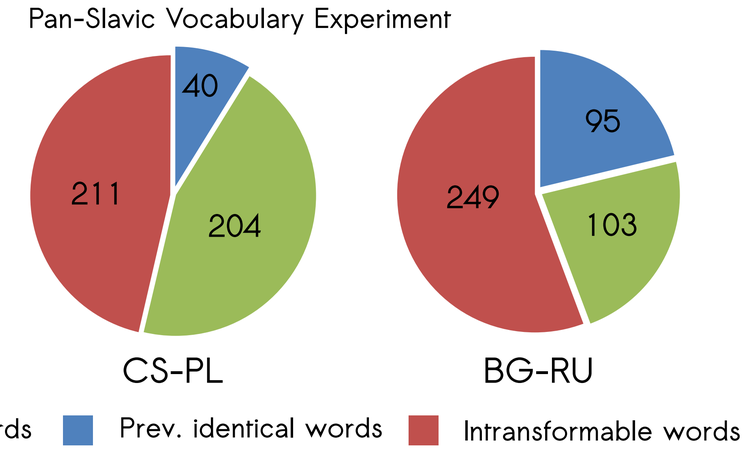
http://www.eurocomslav.de/BIN/inhalt.htm

Applicability of Diachronically-Based Rules

1) Orthographically identical words (8.79% in CZ-PL vs. 21.25% in BG-RU),
2) Application of transformation rules on remaining word pairs: (91.21% vs. 79.75%), but
3) not all word pairs could be covered by rules: morphological differences

→ will be explored in next project phase
→ rules also tested on other word sets (internationalisms)

Swadesh lists with wider vocabulary range/ higher (non-)cognate rates



Summary

Goal: identify mechanisms by which languages en- and decode information
Ideas: - surprisal of language models correlates with intelligibility
- adapt N-gram LMs for cross-language use via latent space and similarity
- analyse information-theoretical results with linguistic knowledge

Next Steps

Linguistically:
Lexis: "false friends" and closed word classes
Morphology: correspondences in grammar
Syntax: word order, complexity of constructions

Information-Theoretically:
Suitable model classes
Most informative features
Inter/intra-language patterns

Contact: C4-all@lsv.uni-saarland.de Web: www.sfb1102.uni-saarland.de/?page_id=296