The Measurement of Mutual Intelligibility between West-Slavic Languages

Perlová Voda, September 2018

Mutual Intelligibility (MI) → Semicommunication

Rozumíte mi dobře?

Len pokiaľ chcem.

MI languages & factors of MI

- Danish – Norwegian – Swedish
- Afrikaans – Frisian – Dutch
- Faroese – Icelandic
- Croatian – Serbian – Slovenian
- Belarusian – Russian – Ukrainian
- Italian – Spanish – Portuguese
- Turkish – Azerbaijani
- ...

Diagram showing categories:
- Attitude and will to cooperate
- Linguistic distance
- Language experience
Research objectives

- Overall mutual intelligibility between West-Slavic languages
- Asymmetry of mutual intelligibility between West-Slavic languages
- Mutual intelligibility of content and function words
- Mutual intelligibility of various styles of material (stylistics)

- Differences between spoken and written forms of West-Slavic languages in all above mentioned areas
Related works

• Dialectometry:
  • (2007) MOBERG J., GOOSKENS Ch., NERBONNE J., VAILLETTE N.

• Sociolinguistics research:
  • (2016) GOLUBOVIC, J.
Method

- Levenshtein distance & Conditional entropy
- Inspired by psycholinguistics idea about process of semicommunication

Rozumíte mi dobře?

Rozumiete mi dobre?

Len pokiaľ chcem.

Jen pokud chci.
Conditional entropy (CE)

- Quantifies the amount of information needed to get the X when Y is given
- Lower entropy = better mutual intelligibility (smaller linguistic distance)
- Allows asymmetrical results (from the definition of CE)

\[
H(X \mid Y) = - \sum_{x \in X, y \in Y} p(x, y) \log_2 (p(x \mid y))
\]

- X ... native language,  x ... native phoneme/grapheme
- Y ... foreign language,  y ... foreign phoneme/grapheme
### CE - example

$$H(X | Y) = - \sum_{x \in X, y \in Y} p(x, y) \log_2 (p(x | y))$$

|   | r | c | z | u | m | i | t | e | m | i | d | c | b | r | e |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CS |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| SK |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| p(CS|SK) | .50 | 1 | 1 | 1 | .67 | 1 | 1 | .75 | .67 | 1 | 1 | 1 | .50 | .75 |   |   |
| p(SK|CS) | 1 | 1 | 1 | .50 | 1 | 1 | .50 | 1 | 1 | .50 | 1 | 1 | 1 | 1 | 1 |   |

Asymmetries: \{ r, r, \varepsilon, \varepsilon, \iota, \upsilon, ... \}

<table>
<thead>
<tr>
<th></th>
<th>j</th>
<th>\varepsilon</th>
<th>n</th>
<th>p</th>
<th>c</th>
<th>k</th>
<th>\upsilon</th>
<th>t</th>
<th>x</th>
<th>ts</th>
<th>i</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p(CS</td>
<td>SK)</td>
<td>1</td>
<td>.75</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.25</td>
<td>.33</td>
</tr>
<tr>
<td>p(SK</td>
<td>CS)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.50</td>
<td>.50</td>
<td>1</td>
<td>1</td>
<td>.50</td>
</tr>
</tbody>
</table>
Material

- corpora: InterCorp v9 2016 (ČNK)
- subcorpora: Acquis, Europarl, Core, Subtitles
- loaded from: KonText v0.9.3
- translations: Treq v1.1
- sample size: 2 000 most frequently used words
- transcription: IPA (semi-automatic)
Results: Overall MI (all w.)

- MI on phonetic layer ≈ MI on graphemic layer
- CS-SK < SK-PL < CS-PL

↓ Agree with socioling. research

The most MI for:
- CS-SK = Europarl, Acquis;
- CS-PL = Core, Subtitles;
- SK-PL = Core, Subtitles.

Subtitles ≈ middle of groups
Results: Asymmetry of MI between CS-SK (all w.)

Phonetic layer:

SK > CS (RA = 0.068)

Graphemic layer:

CS > SK (RA = 0.029)

Agree with socioling. research, except graph.

Same side for all subcorpora across layers

\[
RA = \frac{H(CS|SK) - H(SK|CS)}{H(CS|SK)}
\]
Results: Asymmetry of MI between CS-PL (all w.)

Phonetic layer:
\[ PL > CS \text{ (RA} = 0.017) \]

Graphemic layer:
\[ PL > CS \text{ (RA} = 0.026) \]

Agree with socioling. research, except phon.

Same side for all sub corpora across layers
Results: Asymmetry of MI between SK-PL (all w.)

Phonetic layer:
SK > PL (RA = 0.019)

Graphemic layer:
PL > SK (RA = 0.025)

 Agree with socioling. research

Same side for all subcorpora across layers
Results: MI & asym. on phon. layer (all, content, function w.)

Same trend at the graphemic layer...
Future: What could be improved?

• Data
  – Usable parallel corpora aligned word-by-word

• Levenshtein method
  – CE without aligning by Lev. distance were not so different
  – Need to add constraints or additional rules for aligning
    example:  
    CS: x aː p œ #  
    SK: x aː p ε m
    CS: x aː p # œ
    SK: x aː p ε m

• Conditional entropy
  – Statistical validation of this method (realized only for Scandinavian languages)
Thank you.
References