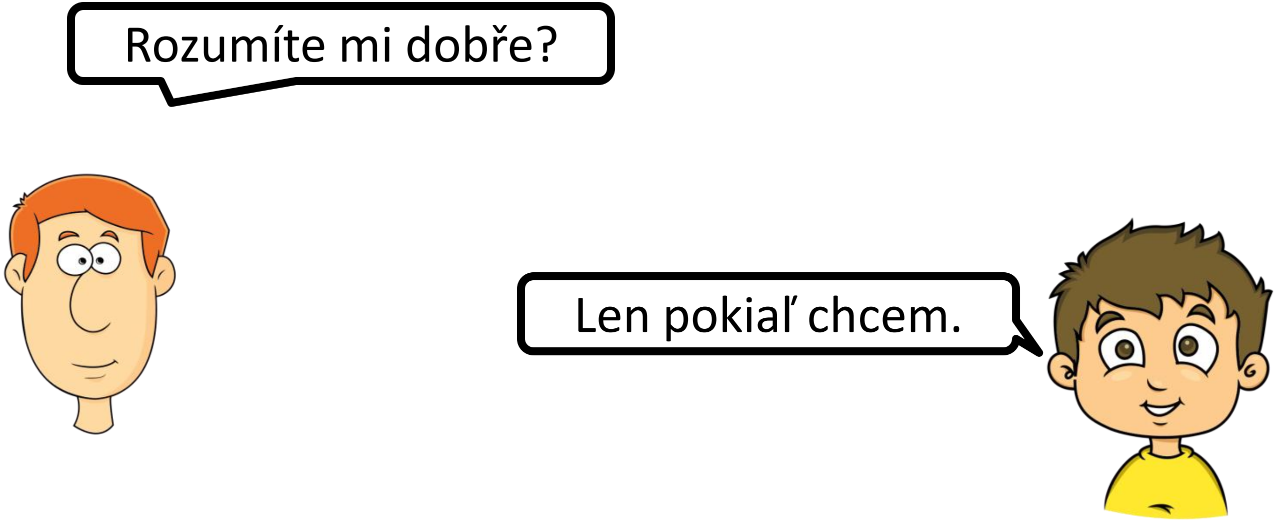
The Measurement of Mutual Intelligibility between West-Slavic Languages

Perlová Voda, September 2018

KYJÁNEK, L. & HAVIGER, J. (2018, June 11): The Measurement of Mutual Intelligibility between West-Slavic Languages. *Journal of Quantitative Linguistics*. DOI: 10.1080/09296174.2018.1464546

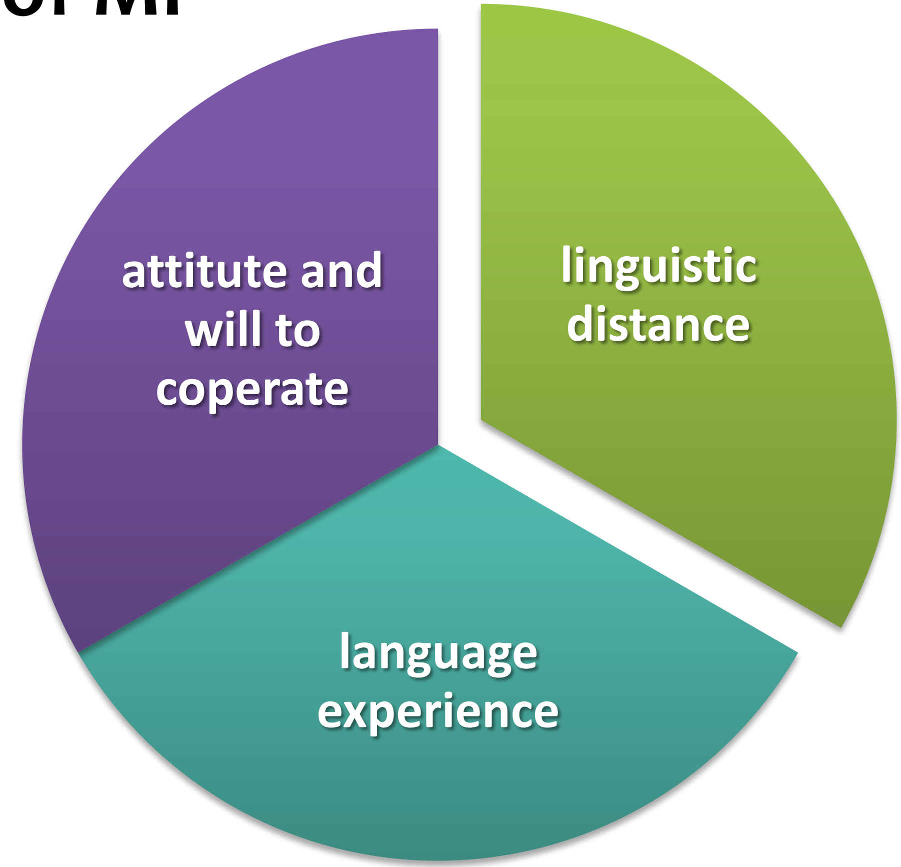
Mutual Intelligibility (MI) → Semicommunication



HAUGEN, E. (1966). Semicommunication: The Language Gap in Scandinavia.

MI languages & factors of MI

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



attitute and will to coperate

linguistic distance

language experience

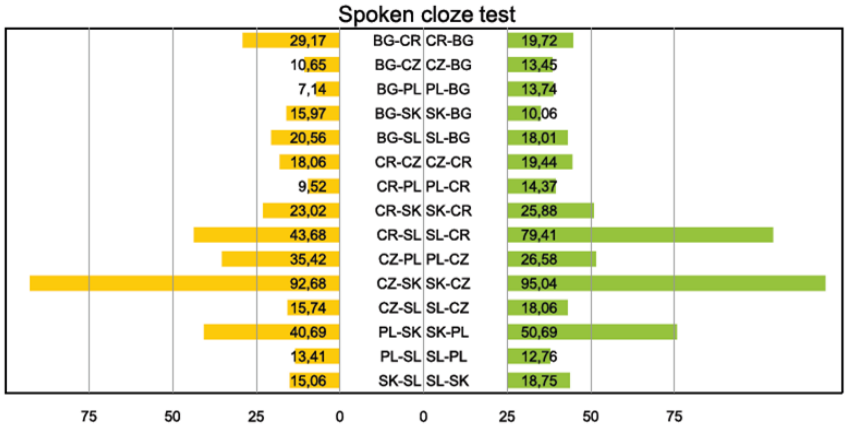
Research objectives

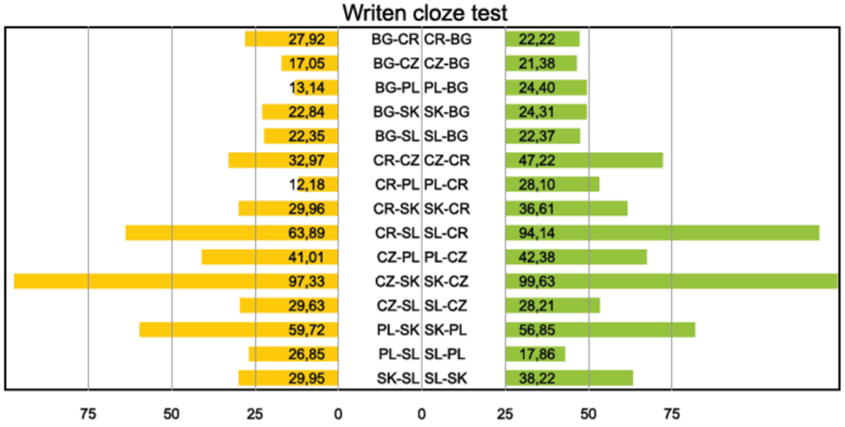
- Overall mutual intelligibility between West-Slavic languages
- Asymmetry of mutual intelligibility between West-Slavic languages
- Mutual intelligibility of content and function words
- Mutual intelligibility of various styles of material (stylistics)

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

Related works

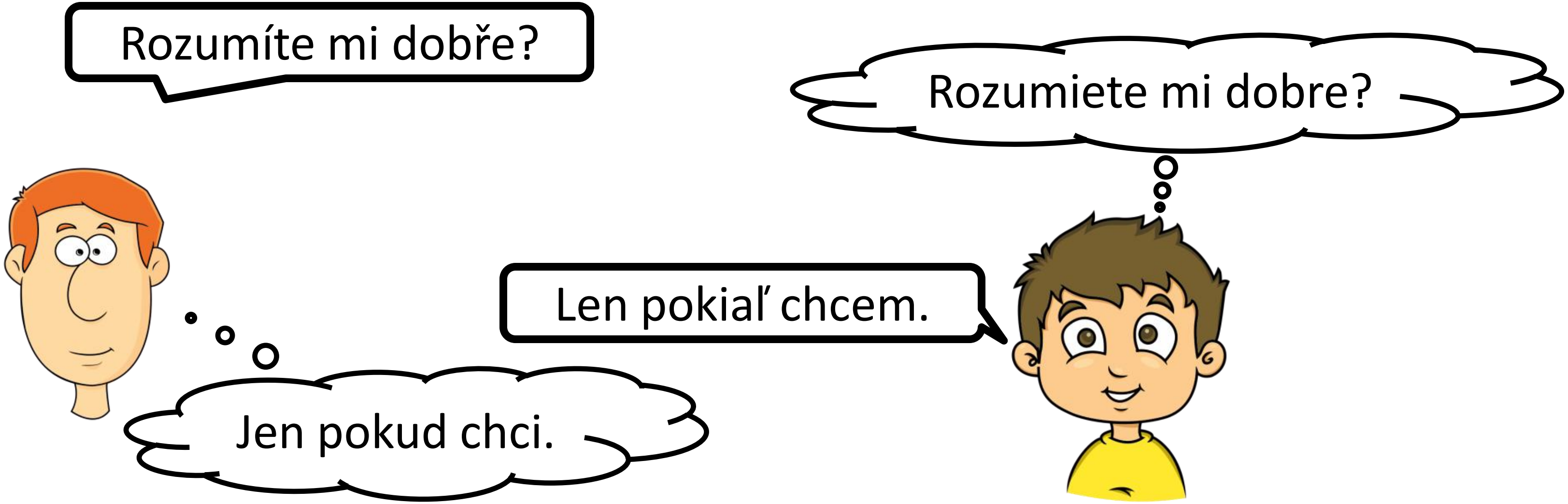
- Dialectometry:
  - (2007) MOBERG J., GOOSKENS Ch., NERBONNE J., VAILLETTE N.
- Sociolinguistics research:
  - (2016) GOLUBOVIĆ, J.
  - (2012), (2009), (2000), (1987)





Method

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



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)

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

- X … native language, x … native phoneme/grapheme

  Y … foreign language, y … foreign phoneme/grapheme

CE - example

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

CS	r	ɔ	z	ʊ	m	iː	t	ɛ		m	ɪ		d	ɔ	b	ɼ	ɛ
SK	r	ɔ	z	ʊ	m	i̯ɛ	t	ɛ		m	i		d	ɔ	b	r	ɛ
p(CS|SK)	.50	1	1	1	.67	1	1	.75		.67	1		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, ɼ }, ɛ { ɛ, ɪ }, ʊ { i̯a, ʊ }, …

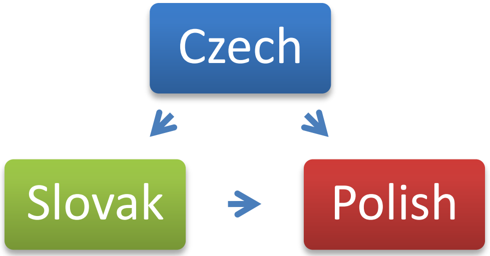
CS	j	ɛ	n		p	ɔ	k	ʊ	t		x	ʦ	ɪ	#
SK	ɭ	ɛ	n		p	ɔ	k	i̯a	ʎ		x	ʦ	ɛ	m
p(CS|SK)	1	.75	1		1	1	1	1	1		1	1	.25	.33
p(SK|CS)	1	1	1		1	1	1	.50	.50		1	1	.50	1

Material

- corpora: **InterCorp v9 2016 (ČNK)**
- subcorpora: **Acquis, Europarl, Core, Subtitles**

- loaded from: **KonText v0.9.3**
- translations: **Treq v1.1**

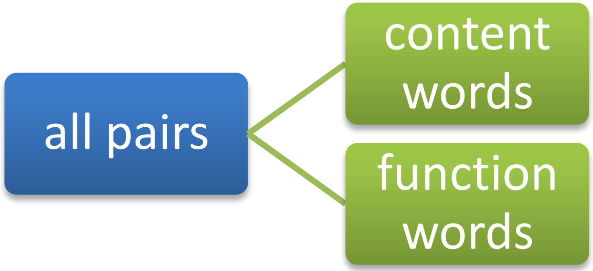
- sample size: **2 000 most frequently used words**
- transcription: **IPA (semi-automatic)**



Czech

Slovak

Polish

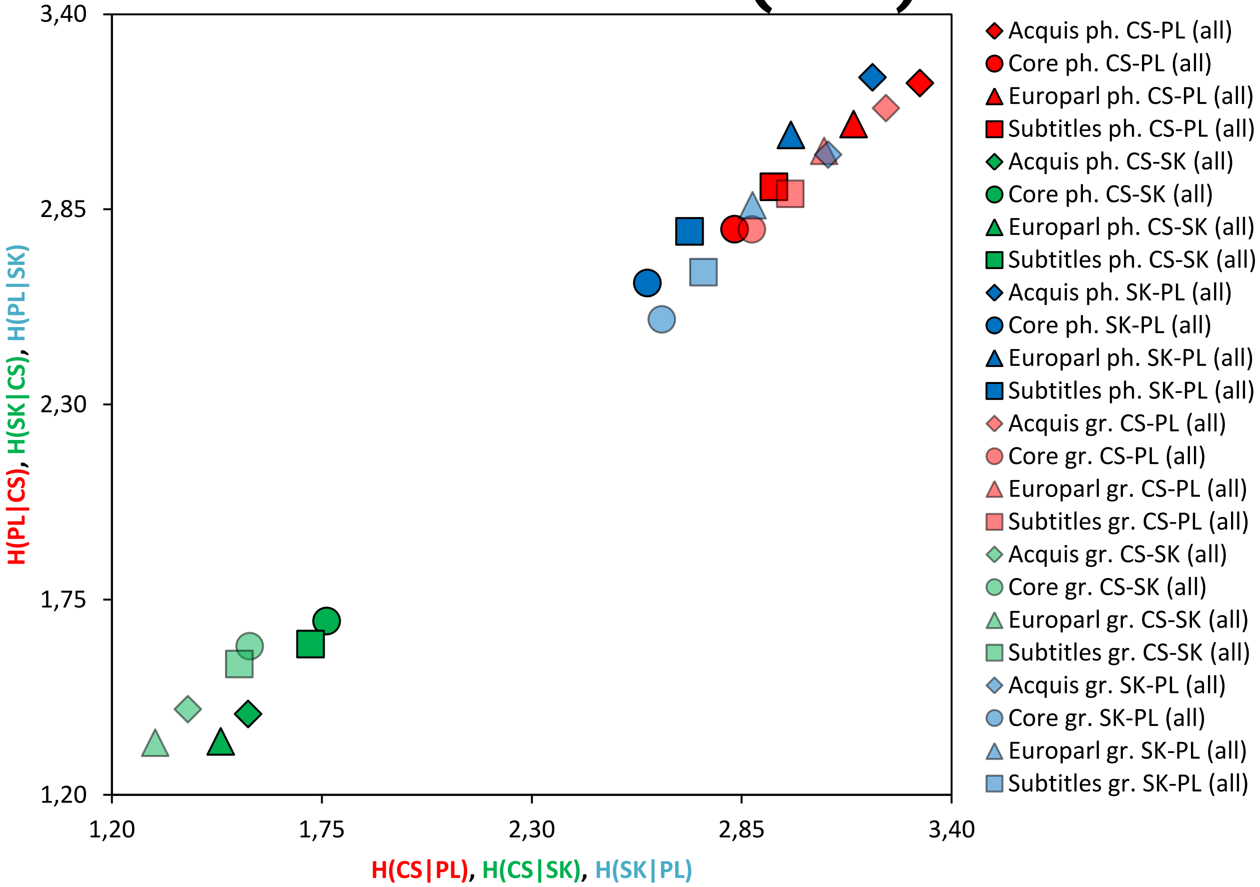


all pairs

content words

function words

Results: Overal MI (all w.)



MI on phonetic layer ≈ MI on graphemic layer

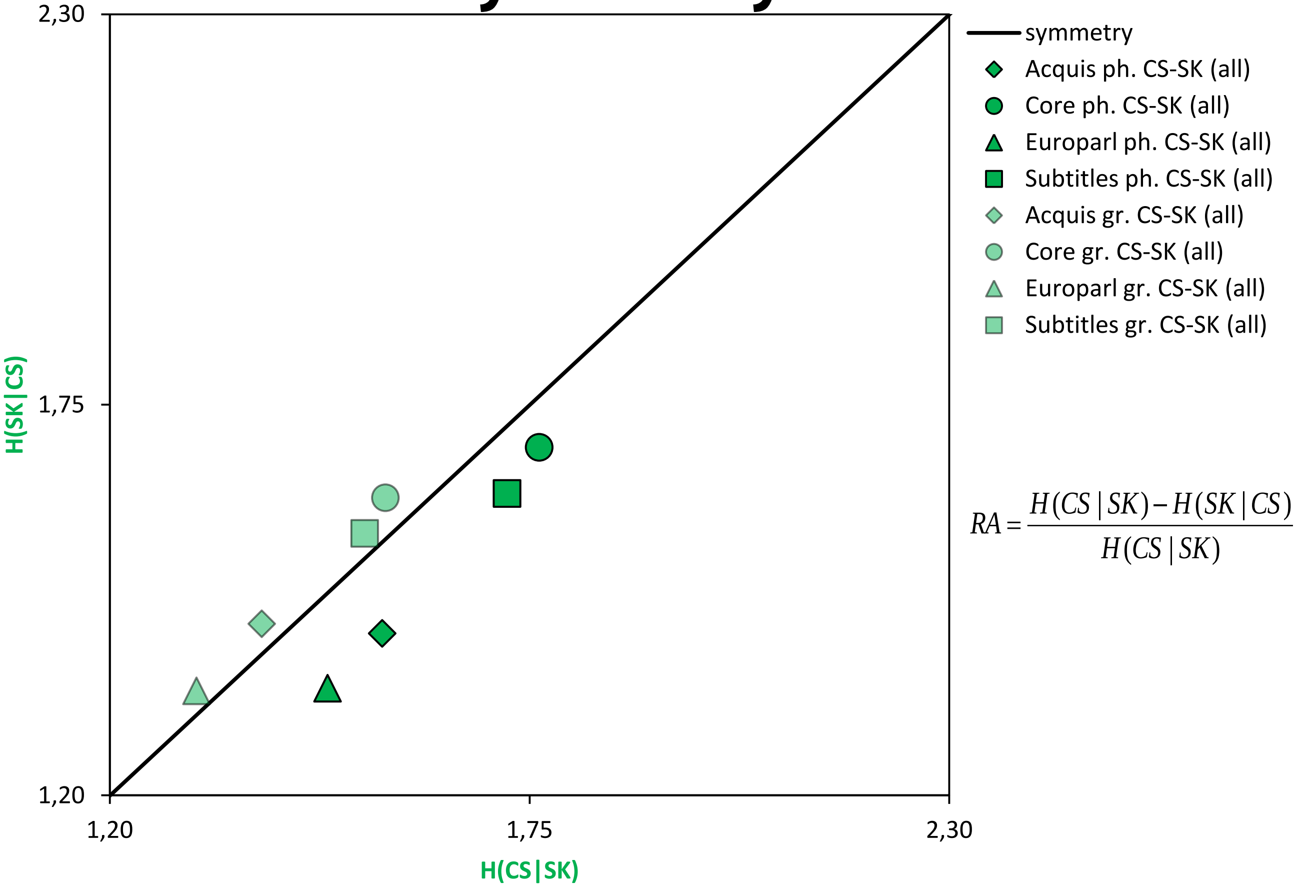
***CS-SK* < SK-PL < CS-PL**

↳ 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.)



$$RA = \frac{H(CS\,|\,SK) - H(SK\,|\,CS)}{H(CS\,|\,SK)}$$

Phonetic layer:

SK > CS (RA = 0,068)

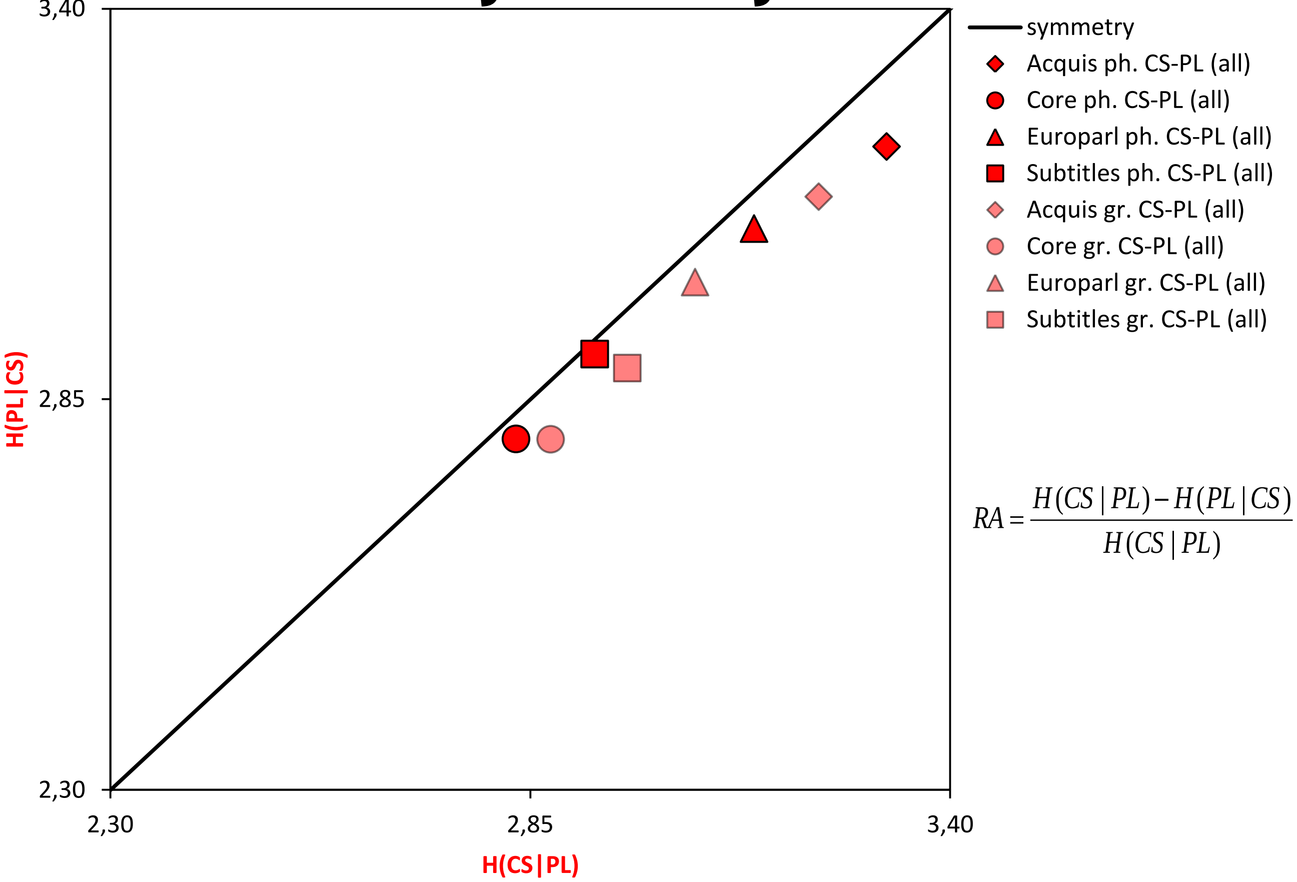
Graphemic layer:

CS > SK (RA = 0,029)

↳ Agree with socioling. research, except graph.

Same side for all subcorpora across layers

Results: Asymmetry of MI between CS-PL (all w.)



$$RA = \frac{H(CS|PL) - H(PL|CS)}{H(CS|PL)}$$

Phonetic layer:

PL > CS (RA = 0,017)

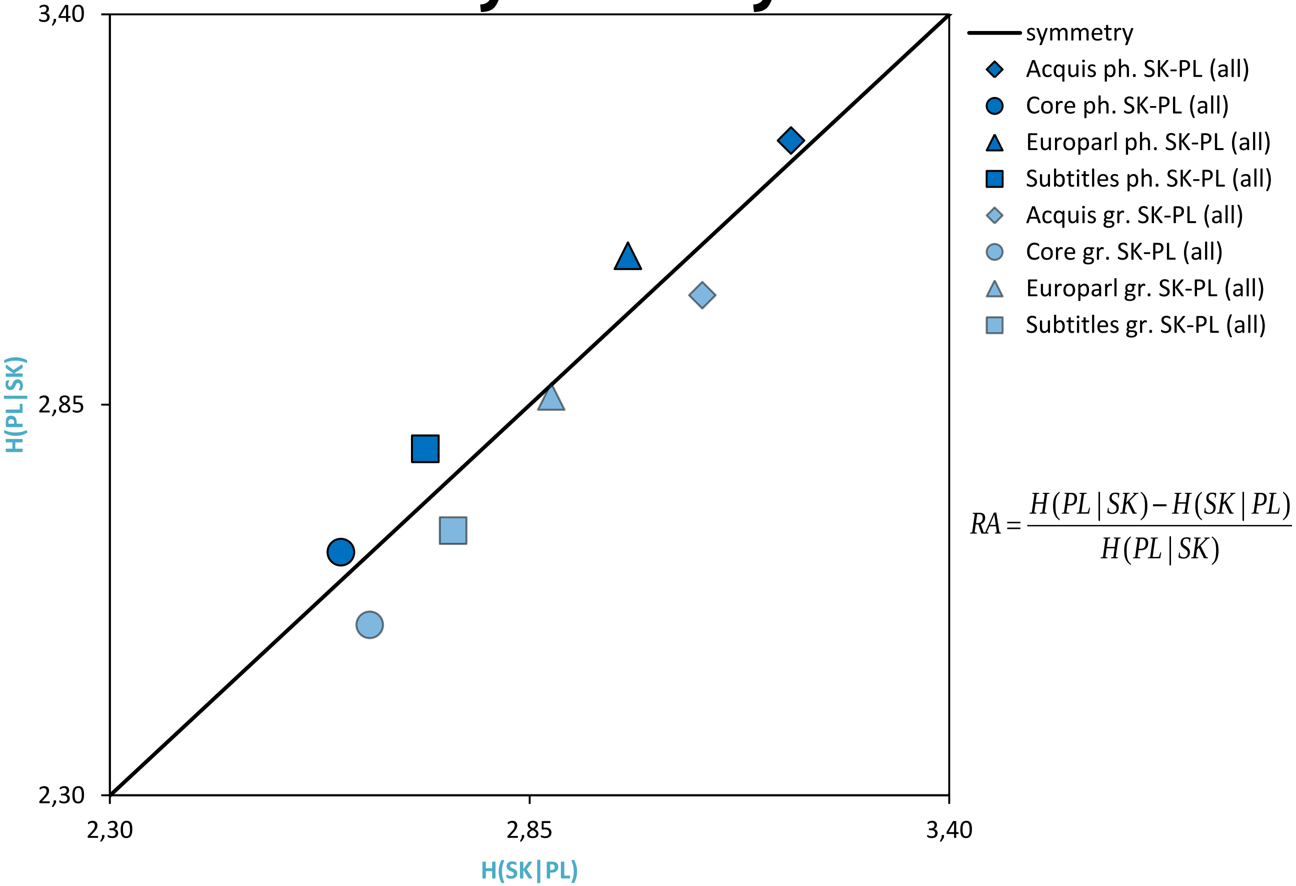
Graphemic layer:

PL > CS (RA = 0,026)

↳ Agree with socioling. research, except phon.

Same side for all subcorpora across layers

Results: Asymmetry of MI between SK-PL (all w.)



$$RA = \frac{H(PL|SK) - H(SK|PL)}{H(PL|SK)}$$

Phonetic layer:

SK > PL (RA = 0,019)

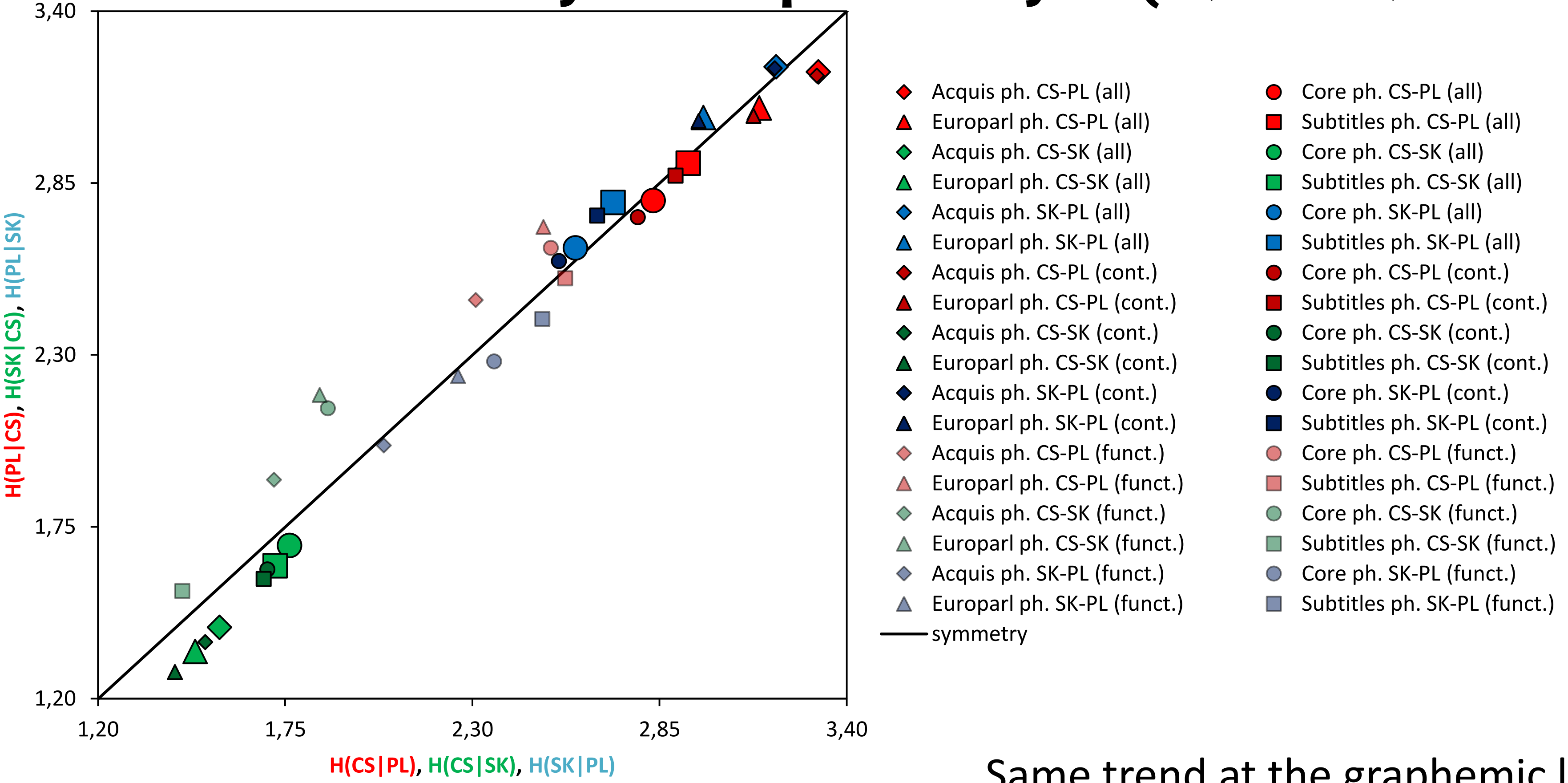
Graphemic layer:

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 ʊ #	CS: x aː p # ʊ
	SK: x aː p ɛ m	SK: x aː p ɛ m

- Conditional entropy
  - Statistical validation of this method (realized only for Scandinavian languages)

Thank you.

References

- HAUGEN, Einar. (1966). Semicommunication: The Language Gap in Scandinavia*.* *Sociological Inquiry*, **36**: 280-297.
- GOLUBOVIĆ, Jelena. (2016). Mutual intelligibility in the Slavic language area [Groningen]: University of Groningen
- BUDOVIČOVÁ, Viera. (1987). Semikomunikácia jako lingvistický problém. *Studia Academica Slovaca*, **16**: 49-66.
- MUSILOVÁ, Květoslava. (2000). *Město a jeho jazyk*. Bratislava: Vydavatel´stvo slovenskej akadémie vied.
- NEKVAPIL, Jiří, SLOBODA, Marián & WAGNER, Petr. (2009). *Mnohojazyčnost v České republice*. Praha: Nakladatelství Lidové noviny.
- DOBROTOVÁ, Ivana, MUSILOVÁ, Květoslava. (2012). Z dějin kontaktu dvou blízkých slovanských jazyků. *Bohemistyka*, 2012(**1**): 35-60.
- Heeringa, Wilbert Jan. (2004). Measuring Dialect Pronunciation Differences using Levenshtein Distance. [Groningen]: University of Groningen
- MOBERG Jens, GOOSKENS Charlotte, NERBONNE John, VAILLETTE Nathan. (2007). Condition Entropy Measures Intelligibility among Related Languages. In: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste and Frank Van Eynde (eds.). *Computational Linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting*. Utrecht: LOT, 51-66.