

Měření jazykové vzdálenosti mezi západoslovanskými jazyky

Lukáš Kyjánek



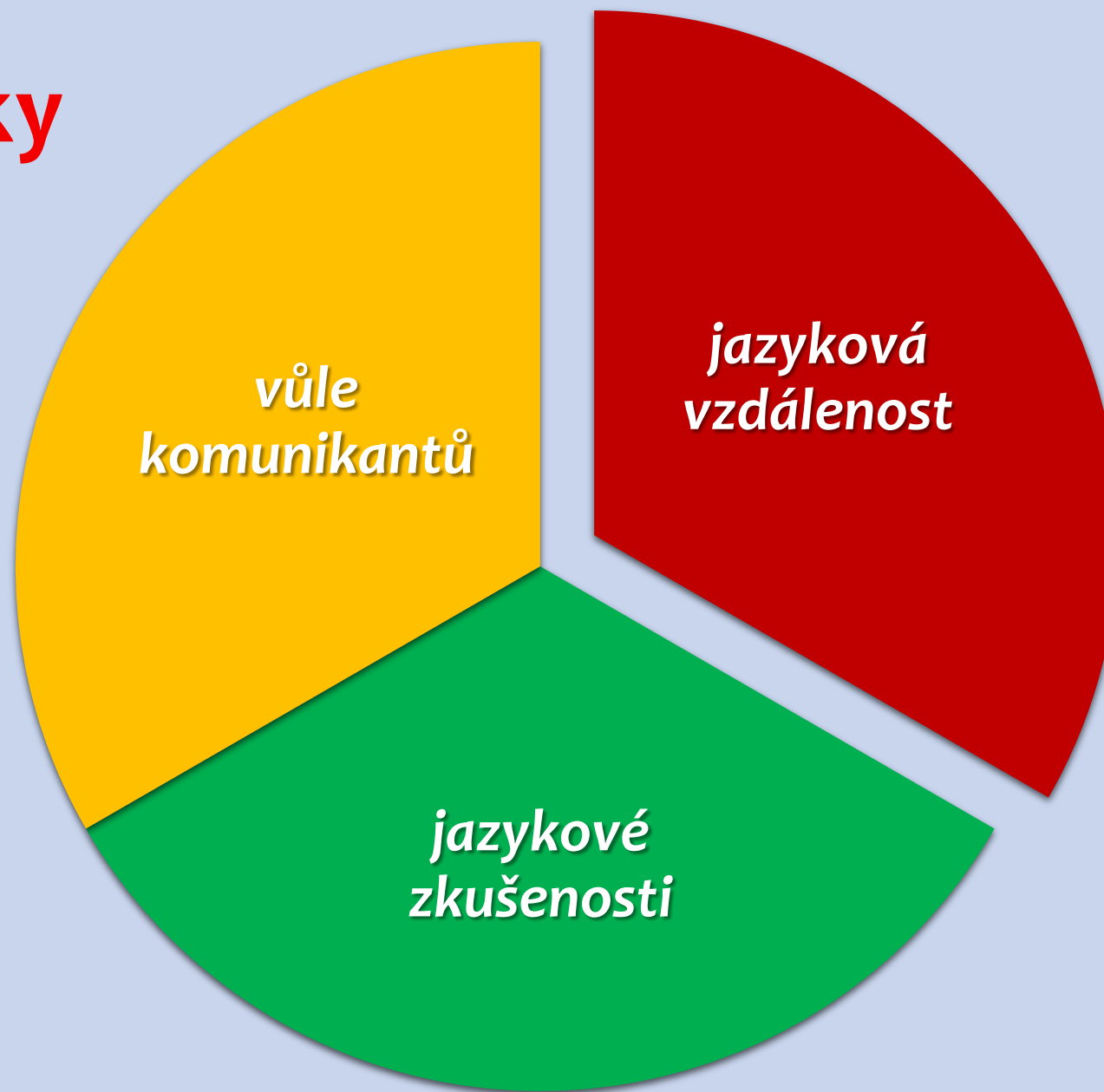
Univerzita Hradec Králové

Semikomunikace

Einar HAUGEN, 1966.

Semicommunication: The Language Gap in Scandinavia.

Podmínky

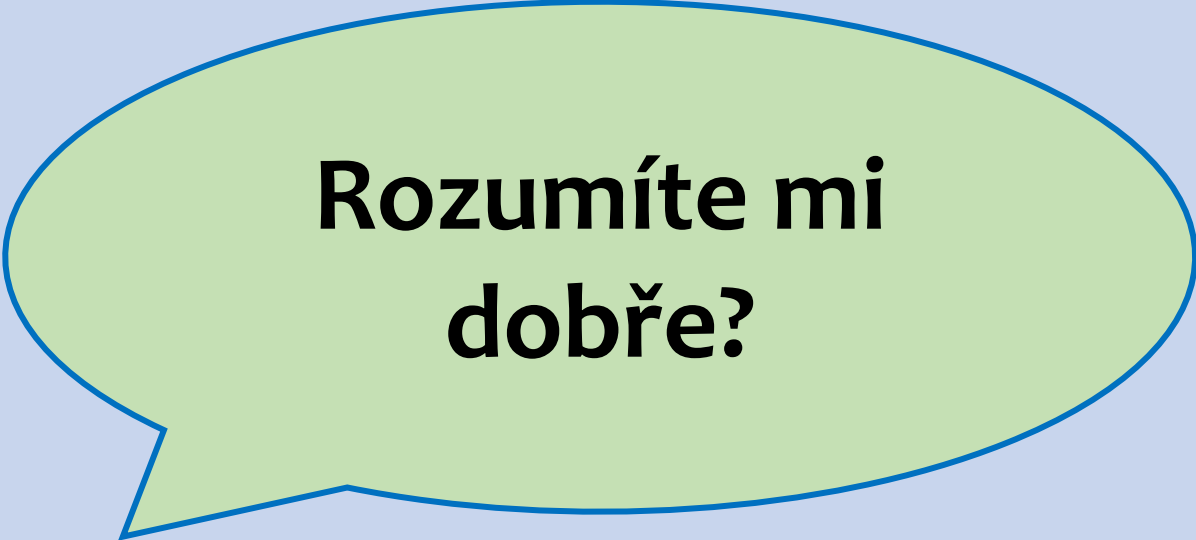


Metody

Jens MOBERG, Charlotte GOOSKENS, John NERBONNE, Nathan VAILLETTE, 2007.

Condition Entropy Measures Intelligibility among Related Languages.

Postup příjemce

A green speech bubble with a blue outline and a tail pointing towards the bottom-left.

**Rozumíte mi
dobře?**

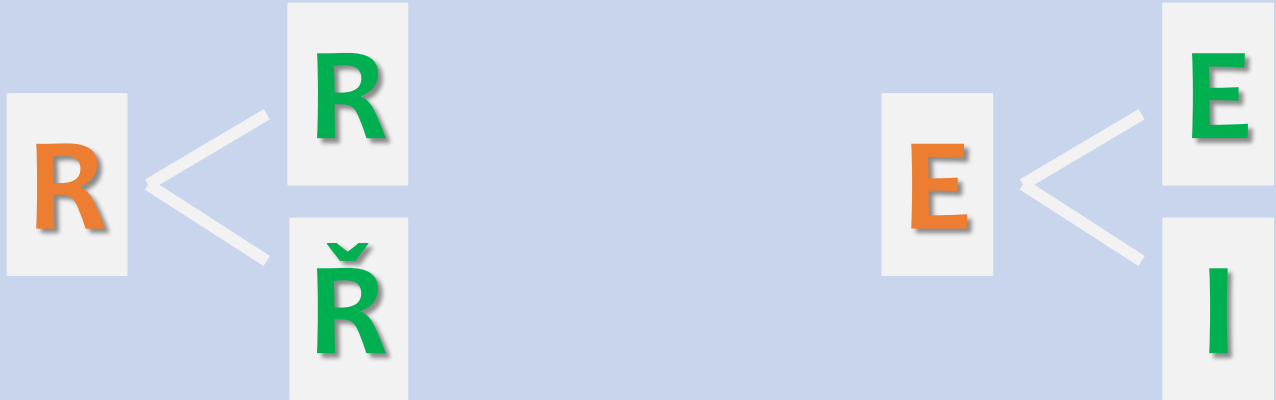
An orange speech bubble with a blue outline and a tail pointing towards the bottom-right.

Len pokiaľ chcem.

CS	R	O	Z	U	M	Í	T	E		M	I		D	O	B	Ř	E	?
SK	R	O	Z	U	M	IE	T	E		M	I		D	O	B	R	E	?

SK	L	E	N		P	O	K	IA	L'		CH	C	E	M	.
CS	J	E	N		P	O	K	U	D		CH	C	I	#	.

CS	R	O	Z	U	M	Í	T	E		M	I		D	O	B	Ř	E	?
SK	R	O	Z	U	M	IE	T	E		M	I		D	O	B	R	E	?



SK	L	E	N		P	O	K	IA	L'		CH	C	E	M	.
CS	J	E	N		P	O	K	U	D		CH	C	I	#	.

Levenštejnova metrika

- Udává minimální počet nutných operací se znaky k tomu, aby byly oba vložené řetězce totožné.
- zarovnávání fonémů

	SK	d	ɔ	b	r	ε
CS	0	1	2	3	4	5
d	1	0	1	2	3	4
ɔ	2	1	0	1	2	3
b	3	2	1	0	1	2
r	4	3	2	1	1	2
ε	5	4	3	2	2	1

CS	d	ɔ	b	r	ε
SK	d	ɔ	b	r	ε

CS	r	ɔ	z	u	m	iː	t	ɛ		m	ɪ		d	ɔ	b	r	ɛ
SK	r	ɔ	z	u	m	ḷ̣ɛ	t	ɛ		m	i		d	ɔ	b	r	ɛ

87,5 %

50 %

80 %

CS	j	ɛ	n		p	ɔ	k	u	d		x	ts	ɪ	#
SK	ḷ	ɛ	n		p	ɔ	k	ḷ̣a	ʎ		x	ts	ɛ	m

67 %

60 %

50 %

65,75 %

Podmíněná entropie

- Vyjadřuje míru nejistoty z pohledu jednotlivých jazyků zvlášť, a proto je schopna postihnout i asymetrické děje.
- výpočty jazykové vzdálenosti a vzájemného porozumění
- Čím je menší hodnota, tím je menší vzdálenost (lepší porozumění).

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

CS	r	ɔ	z	ʊ	m	iː	t	ɛ		m	i		d	ɔ	b	r	ɛ
SK	r	ɔ	z	ʊ	m	ɿɛ	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

$$p(x|y) = \frac{p(x \cap y)}{p(y)} \Rightarrow \text{asymetrie}$$

CS	j	ɛ	n		p	ɔ	k	ʊ	t		x	ts	i	#
SK	ɭ	ɛ	n		p	ɔ	k	ɿa	ʎ		x	ts	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

Materiál



- korpus: **InterCorp v9 2016**
- subkorpus: **Subtitles**

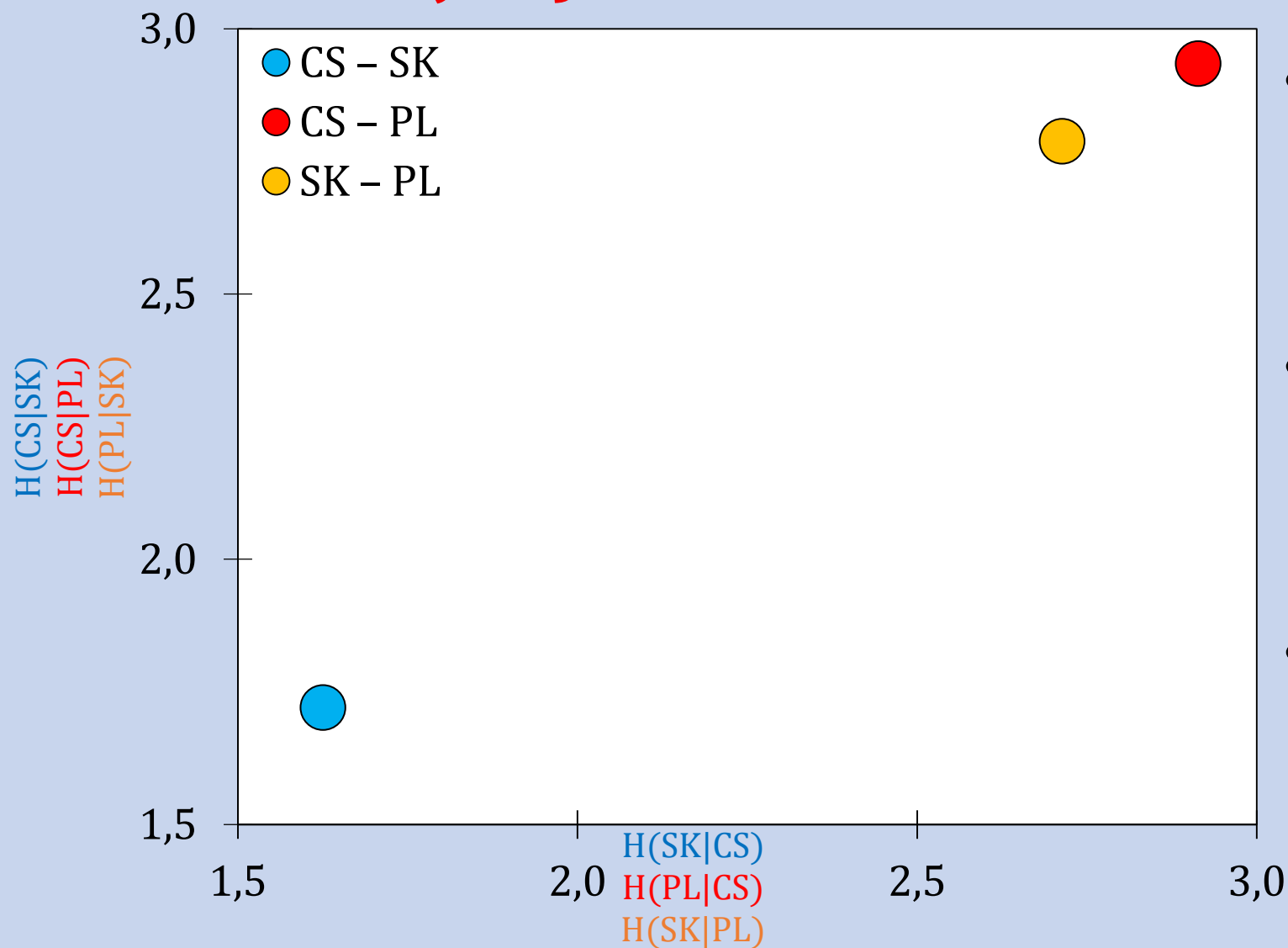
- data: **KonText v0.9.3** (*Vlastní frekvenční distribuce*)
- překlady: **Treq v1.1**

- vzorek: **2 000 nejfrekventovanějších slov** (*word*)
- transkripce: **IPA**

Výsledky výzkumu

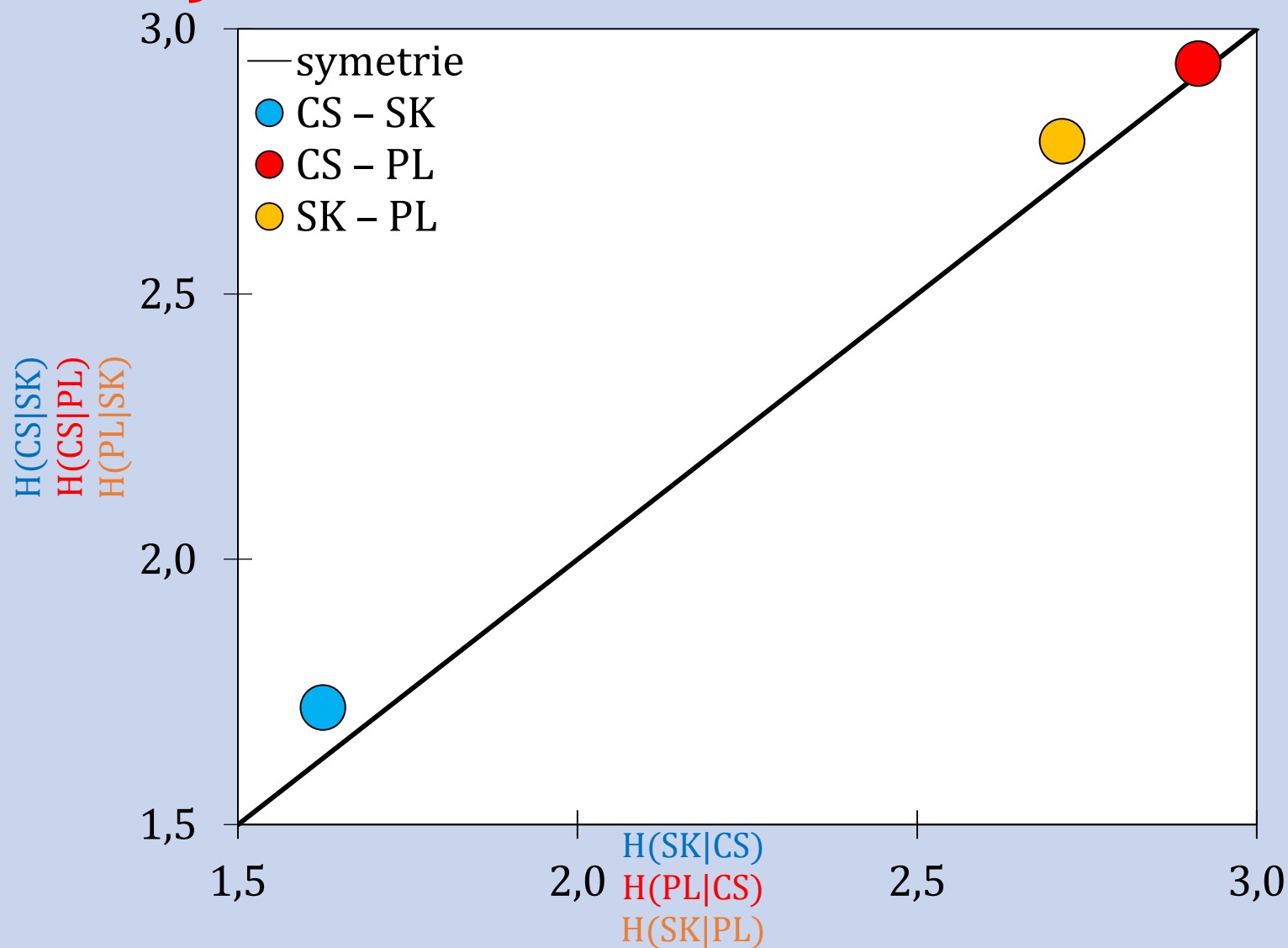
Měření podmíněné entropie mezi češtinou a slovenštinou, češtinou a polštinou, slovenštinou a polštinou.

Celková jazyková vzdálenost



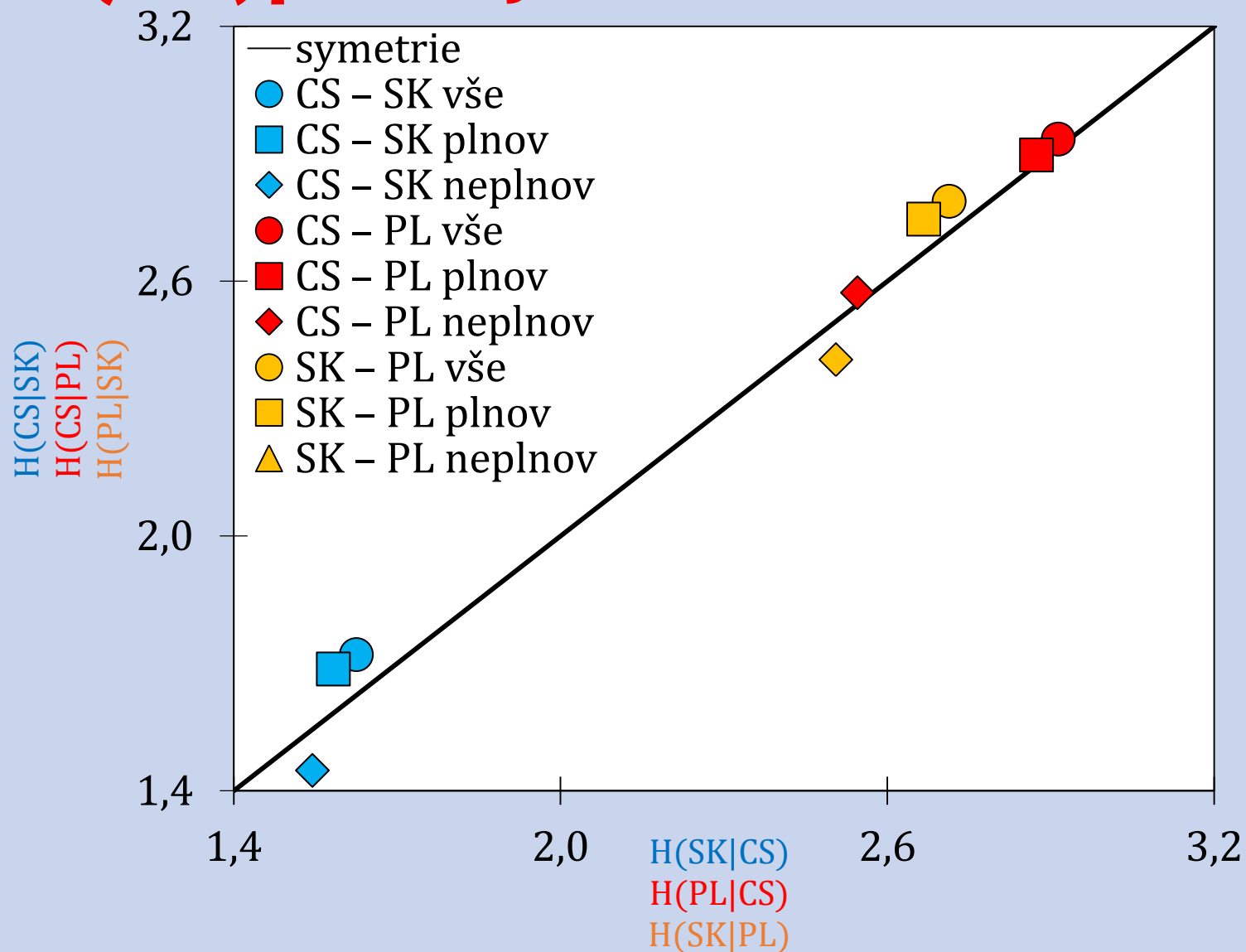
- **Čeština a slovenština mají nejmenší jazykovou vzdálenost.**
- **Slovenština a polština mají vysokou jazykovou vzdálenost.**
- **Čeština a polština mají největší jazykovou vzdálenost.**

Asymetrie



- *Slováci rozumí češtině lépe než Češi slovenštině.*
- *Poláci rozumí češtině lépe než Češi polštině.*
- *Slováci rozumí polštině lépe než Poláci slovenštině.*

(Ne)plnovýznamovost



- *Plnovýznamová slova mají hodnoty blízké všem slovům.*
- *Neplnovýznamová slova jsou si méně foneticky vzdálená než plnovýznamová slova.*
- *Neplnovýznamová slova stojí na opačné straně asymetrie než plnovýznamová slova.*

