Tania Avgustinova, Universität des Saarlandes

INCOMSLAV

SFB 1102

Gegenseitige Verstehbarkeit / Verständlichkeit und Surprisal in slavischer Interkomprehension: empirische Basis und linguistische Modellierung

Mutual Intelligibility and Surprisal in Slavic Intercomprehension: Empirical Base and Linguistic Modelling

SFB in Saarbrücken

INCOMSLAV
SFB 1102

Sprachgebrauch:
Die Sprache bietet eine Vielzahl von Optionen zum Codieren einer Nachricht.

Sprachvariation:
Variation ist eine inhärente Eigenschaft des Sprachsystems

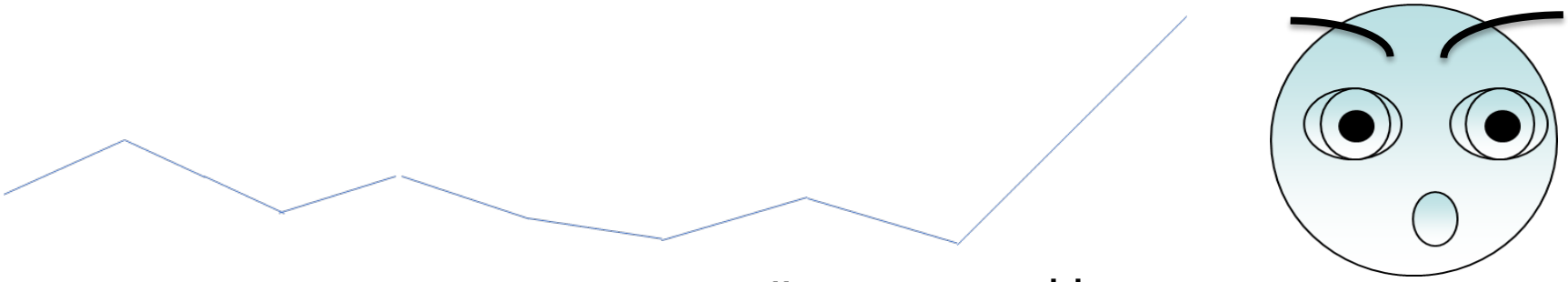
- Zentale Hypothesen:
  - Die Sprachverarbeitung beruht auf der **Vorhersehbarkeit im Kontext**
  - Die kontextuell bestimmte Vorhersagbarkeit kann durch den Shannons Begriff der **Information** angemessen indiziert werden

Information Density
Surprisal

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

Illustration von Surprisal

Sie ging in den Laden und kaufte Äpfel und Orangen.



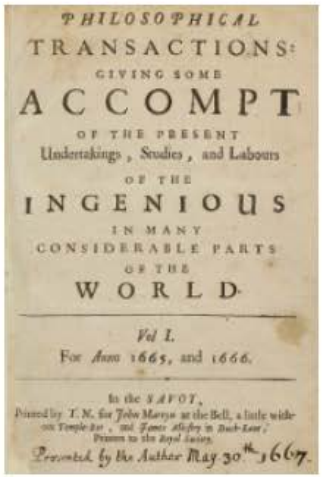
Sie ging in den Laden und kaufte Äpfel und Hexagons.

SFB in Saarbrücken: Forschungsgebiete

INCOMSLAV
SFB 1102

A *Situational Context and World Knowledge*
Brings non-linguistic context into characterizations of surprisal

B *Discourse and Register*
Examines the relation between encoding and information density at the level of text



C *Variation in Linguistic Encoding*
Offers information density explanations for encoding choices across linguistic levels and languages

SFB in Saarbrücken: Forschungsteam

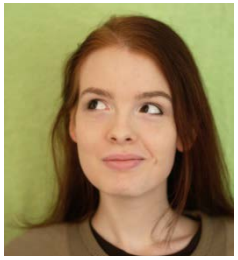
INCOMSLAV
SFB 1102

Bernd Möbius
Phonetics

Dietrich Klakow
Statistical NLP

Roland Marti
Slavic Studies

Tania Avgustinova
Computational & Slavic Linguistics



Fokus dieses Vortrags

- **Gegenseitige Verstehbarkeit** (mutual intelligibility)
  - **Ähnlichkeiten** zwischen Sprachen generieren **Erwartungen** bezüglich linguistischer **Kodierung**
  - Ziel: statistische Evidenz von gegenseitiger Verständlichkeit zu finden
- **Surprisal** (Überraschungswert) als Maß vom Informationsgehalts
  - Umkehrung der Wahrscheinlichkeit: unwahrscheinliche (überraschende) Ereignisse enthalten mehr Informationen
  - Idee: Surprisal von Sprachmodellen korreliert mit der Verständlichkeit
- **Slavische Interkomprehension**
  - sprachübergreifende Toleranz gegenüber ungewohnter Kodierung
  - stützt sich auf sinnvolle sprachliche Einheiten
    - verminderte Verständlichkeit (durch fehlende Einheiten)
    - Verwirrung (durch falsche Erkennung von Einheiten)

Überblick

INCOMSLAV
SFB 1102

Empirische Basis		
strukturierte Daten	**„Big Data" – unstrukturiert**	**experimentelle Daten**
slavistische Expertise: historisch-vergleichend	Korpora; Wortlisten	spontane Interkomprehension

Linguistische Modellierung			
Distanzberechnungen	**statistische Verfahren**		**Informationstheorie**
Levenshtein-Metrik (Editierdistanz)	MDL	N-Gram	Entropie (Ungewissheit); Surprisal (Informationsgehalt)

Verständlichkeitsgrad bei Interkomprehension

	Distanz groß / Ähnlichkeit gering ↓ unverständliche Kodierung	Distanz gering / Ähnlichkeit groß ↓ verständliche Kodierung
Surprisal hoch ↓ unerwartete Info	am schwierigsten ↓ unverständlich	Beeinflussung durch linguistische Kenntnisse zur gegenseitigen Verständlichkeit möglich
Surprisal niedrig ↓ erwartete Info	Verständnis kann durch außersprachliches Wissen unterstützt werden	trivial ↓ voll verständlich

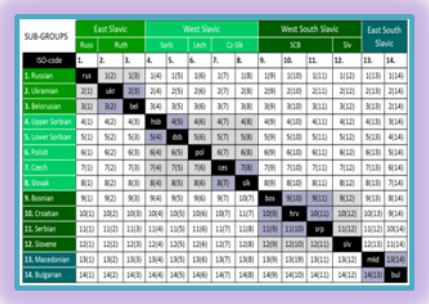
- Weitere Faktoren:
  - erhöhte Schwierigkeit: Täuschung durch Verwandtschaft (falsche Freunde)
  - erleichterte Verarbeitung: Erwartung, überrascht zu werden

Forschungsinfrastruktur

INCOMSLAV
SFB 1102

linguistic encoding

Slavische Interkomprehensionsmatrix



SUB-GROUPS	East Slavic			West Slavic				West South Slavic				East South Slavic	
	Russ	Ruth		Sorb	Lech	Cz-Sk		SCB			Slv		
ISO-code	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Russian	rus	1(2)	1(3)	1(4)	1(5)	1(6)	1(7)	1(8)	1(9)	1(10)	1(11)	1(12)	1(13)	1(14)
2. Ukrainian	2(1)	ukr	2(3)	2(4)	2(5)	2(6)	2(7)	2(8)	2(9)	2(10)	2(11)	2(12)	2(13)	2(14)
3. Belorusian	3(1)	3(2)	bel	3(4)	3(5)	3(6)	3(7)	3(8)	3(9)	3(10)	3(11)	3(12)	3(13)	2(14)
4. Upper Sorbian	4(1)	4(2)	4(3)	hsb	4(5)	4(6)	4(7)	4(8)	4(9)	4(10)	4(11)	4(12)	4(13)	3(14)
5. Lower Sorbian	5(1)	5(2)	5(3)	5(4)	dsb	5(6)	5(7)	5(8)	5(9)	5(10)	5(11)	5(12)	5(13)	4(14)
6. Polish	6(1)	6(2)	6(3)	6(4)	6(5)	pol	6(7)	6(8)	6(9)	6(10)	6(11)	6(12)	6(13)	5(14)
7. Czech	7(1)	7(2)	7(3)	7(4)	7(5)	7(6)	ces	7(8)	7(9)	7(10)	7(11)	7(12)	7(13)	6(14)
8. Slovak	8(1)	8(2)	8(3)	8(4)	8(5)	8(6)	8(7)	slk	8(9)	8(10)	8(11)	8(12)	8(13)	7(14)
9. Bosnian	9(1)	9(2)	9(3)	9(4)	9(5)	9(6)	9(7)	10(7)	bos	9(10)	9(11)	9(12)	9(13)	8(14)
10. Croatian	10(1)	10(2)	10(3)	10(4)	10(5)	10(6)	10(7)	11(7)	10(9)	hrv	10(11)	10(12)	10(13)	9(14)
11. Serbian	11(1)	11(2)	11(3)	11(4)	11(5)	11(6)	11(7)	11(8)	11(9)	11(10)	srp	11(12)	11(12)	10(14)
12. Slovene	12(1)	12(2)	12(3)	12(4)	12(5)	12(6)	12(7)	12(8)	12(9)	12(10)	12(11)	slv	12(13)	11(14)
13. Macedonian	13(1)	13(2)	13(3)	13(4)	13(5)	13(6)	13(7)	13(8)	13(9)	13(10)	13(11)	13(12)	mkd	13(14)
14. Bulgarian	14(1)	14(2)	14(3)	14(4)	14(5)	14(6)	14(7)	14(8)	14(9)	14(10)	14(11)	14(12)	14(13)	bul

modelling

experiments

Forschungsinfrastruktur

INCOMSLAV
SFB 1102

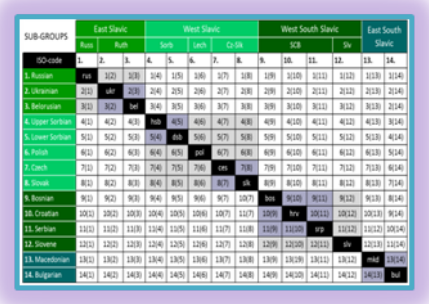
linguistische Phänomene
ORTH, LEX, GRAM

BG-RU (kyrilische Schrift) & PI-CS (lateinsiche Schrift)

linguistic determinants of intelligibility

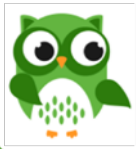
quantitative models of surprisal

Slavische Interkomprehensionsmatrix



test material preparation; hypotheses

observation of intercom-prehension



Vorhersagbarkeit im Kontext
LM GUI: Training & Scoring auf Korpora

Distanzberechnungstool
- Levenshteindistanz (LD, symmetrisch)
- Bedingte Entropie (BE, asymmetrisch)
- Adaptationssurprisal (AS, asymmetrisch)
- Visualisierung

predict human performance

validate surprisal models

Korrelationen

Web-basierte Infrastruktur
Online-Experimenten-Portal

Experimente
- einzelne Wörter
- Phrasen (NP)
- Lückentexte
- Sätze

Admin-Panel
- Stimuli-Upload
- Datensammlung
- Visualisierung
- Grundstatistiken

Slavische Interkomprehensionsmatrix

↓ L1 / Lx →	East Slavic			West Slavic					West South Slavic				East South Slavic	
	Russ	Ruth		Sorb		Lech	Cz-Slk		SCB			Slv		
ISO-code	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**	**8.**	**9.**	**10.**	**11.**	**12.**	**13.**	**14.**
1. Russian	rus	1(2)	1(3)	1(4)	1(5)	1(6)	1(7)			1(10)	1(11)	1(12)	1(13)	1(14)
2. Ukrainian	2(1)	ukr	2(3)	2(4)	2(5)	2(6)	2(7)			2(10)				2(14)
3. Belarusian	3(1)	3(2)	bel	3(4)	3(5)	3(6)	3(7)							2(14)
4. Upper Sorbian	4(1)			hsb	4(5)	4(6)	4(7)		4(9)					3(14)
5. Lower Sorbian	5(1)			5(4)	dsb	5(6)	5(7)	5(8)	5(9)					4(14)
6. Polish	6(1)				6(5)	pol	6(7)	6(8)	6(9)				6(13)	5(14)
7. Czech	7(1)			7(4)	7(5)	7(6)	ces	7(8)	7(9)	7(10)	7(11)	7(12)		6(14)
8. Slovak	8(1)	8(2)	8(3)	8(4)	8(5)	8(6)	8(7)	slk	8(9)	8(10)	8(11)	8(12)		
9. Bosnian	9(1)	9(2)				9(6)	9(7)	10(7)	bos	9(10)	9(11)			
10. Croatian	10(1)	10(							10(9)	hrv	10(11)	10(12)		
11. Serbian	11(1)	1								11(10)	srp	11(12)	11(12)	10(14)
12. Slovene	12(1)	1							12(9)	12(10)	12(11)	slv	12(13)	11(14)
13. Macedonian	13(1)					13(6)			13(9)	13(19)	13(11)	13(12)	mkd	13(14)
14. Bulgarian	14(1)	14(2)	14(3)	14(4)	14(5)	14(6)	14(7)	14(8)	14(9)	14(10)	14(11)	14(12)	14(13)	bul

Czech through Polish

How can a Russian understand Bulgarian?

Polish through Czech

Croatian view on Serbian

How can a Bulgarian understand Russian?

Serbian view on Croatian

Notation: **L1(Lx)**, wo **L1** = *Dekodierungssprache und* **Lx** = *Stimulussprache*

Polnisch mit tschechischen Augen

- **Nie pieprz Pietrze wieprza pieprzem, bo przepieprzysz wieprza pieprzem!**

rz→ř **sz→š** **ie→e** **w→v**

Don't put pepper, Peter, on the pork, for you'll screw the pork with pepper!

Polnisch mit tschechischen Augen

- **Nie pieprz Pietrze wieprza pieprzem, bo przepieprzysz wieprza pieprzem!**

rz→ř **sz→š** **ie→e** **w→v**

Don't put pepper, Peter, on the pork, for you'll screw the pork with pepper!

Polnisch mit tschechischen Augen

Nie pieprz Pietrze wieprza pieprzem, bo przepieprzysz wieprza pieprzem!

rz→ř sz→š ie→e w→v

Nepepř Petře vepře pepřem – přepepříš vepře pepřem!

Don't put pepper, Peter, on the pork, for you'll screw the pork with pepper!

Polnisch mit tschechischen Augen

INCOMSLAV
SFB 1102

1. **PL**: Nie pieprz Pietrze wieprza pieprzem, bo przepieprzysz wieprza pieprzem!

   CZ: Nepepř Petře vepře pepřem – přepepříš vepře pepřem.

→ Die Regelmäßigkeiten sind transparent.

---

2. **PL**: W Szczebrzeszynie chrząszcz brzmi w trzcinie.

   fʂtʂɛbʐɛˈʂɨɲɛ ˈxʂɔ̃ʂtʂ ˈbʐmi fˈtʂtɕiɲɛ

   CZ: V Štěbřešínu chroust břmí v třtině.

→ Die Regelmäßigkeiten sind weniger transparent: Digraphe? Silbengrenzen?

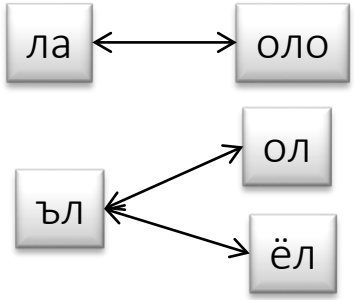
Regelmäßige Korrespondenzen

- **orthographische Korrelate**
  - Slavisches Vokabular (gemeinsames Erbe)
    → historische Korrespondenzregeln
  - Internationalismen (modernes Vokabular)

(RU) В Европейском парламенте

(BG) В Европейския парламент

BG	RU
кон	конь
тяло	тело
море	море
четка	щётка
крава	корова
пред	перед
глава	голова
глас	голос
пълен	полный
жълт	жёлтый
вълк	волк



- **morphologische Korrelate** (Schema vs. Elemente)

(RU) -ом + -е [präp/lok Kasus]

(BG) -ия [def.Art_mask.Adj_kurz]

- **syntactische Muster** (Konstruktionen)

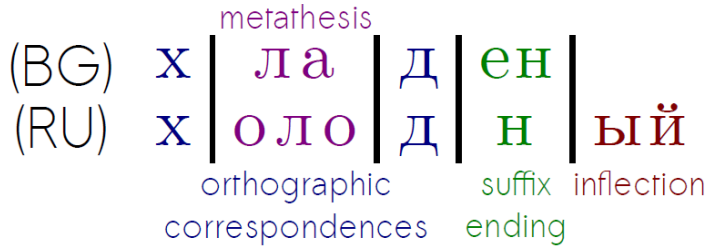
(BG) … с **глава**та надолу …

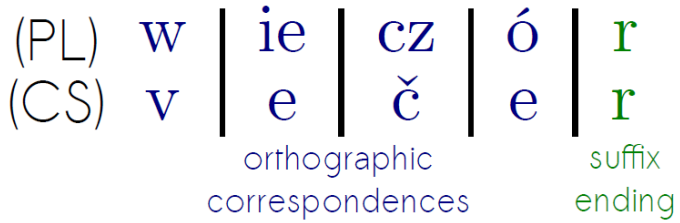
(RU) … вниз **голов**ой …

Quantifizierung linguistischer Ähnlichkeit auf Kognaten

Statistische Entdeckung von Korrelaten

- **Ziel**: Finden orthographischer & morphologischer Korrelate
- **Ressourcen**: CS-PL & RU-BG Kognaten, diachronische Korrespondenzregeln
- **Ergebnis**: Menge potentieller Korrelate (Orthographie, Flexion, Derivation)





- **Sind die entdeckten Korrespondenzen linguistisch sinnvoll?**

  Das Model reproduziert linguistische Regeln und findet fehlende Korrespondenzregeln!

  - 7/10 fehlende orthographische Korrespondenzen gefunden:

    BG:RU (е:э) ✔ (ъ:ø) ✔ (ъ:е) ✔ (ъ:ё) ✔ (ьо:ё) ✔ (п:пп) ✔ (с:сс) ✔ (л:лл) ✘ (р:рр) ✘ (н:нн) ✘

  - 9/10 fehlende Flexionskorresponenzen gefunden:

    BG:RU (#:ет#) ✔ (#:ый#) ✔ (о#:о#) ✔ (а#:а#) ✔ (#:ий#) ✔ (е#:ет#) ✔ (#:ой#) ✔ (и#:ит#) ✔ (е#:ёт#) ✔ (ø:-ит#) ✘

  - fast alle Stamm-und-Endung-Korrelate gefunden

Objektive Ähnlichkeit auf String-Ebene

- Minimum Description Length
  - MDL basiert auf Datenkomprimierung als Indiz für …
  - … Regelmäßigkeit und Komplexität der gemeinsamen Struktur
  - sowohl benutzt als auch produziert parallele Daten

(BG)	м	и	л	ый
(RU)	м	и	л	ый
(PL)	m	i	ł	y
(CS)	m	i	l	ý

(BG)	п	и		я
(RU)	п	и		ть
(PL)	p		i	ć
(CS)	p		í	t

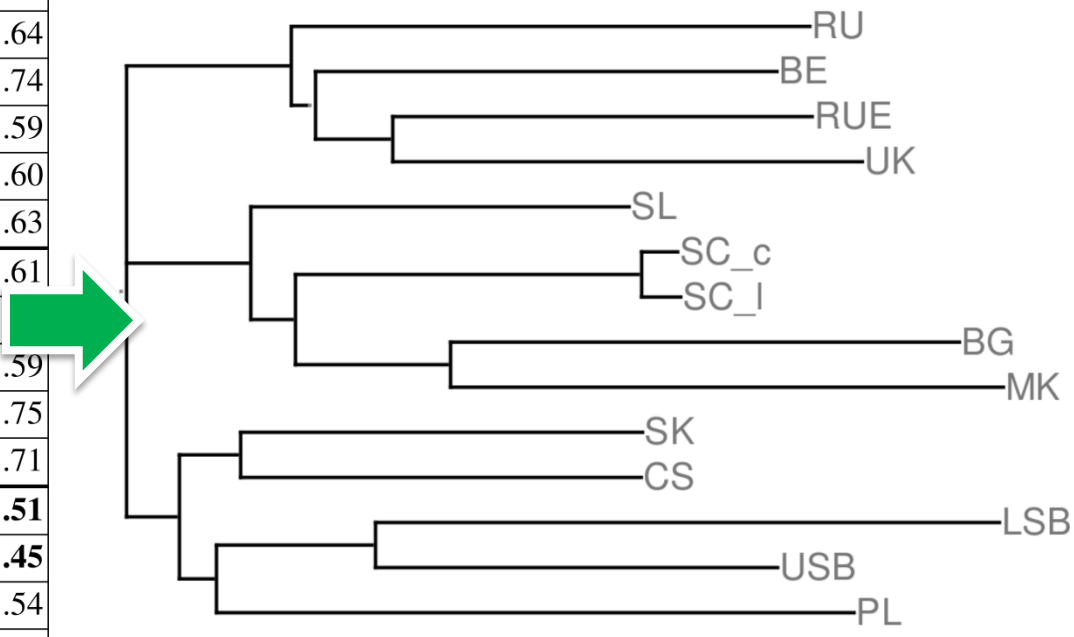
- Was können wir damit machen?

Quantifizierung sprachlicher Ähnlichkeit

INCOMSLAV
SFB 1102

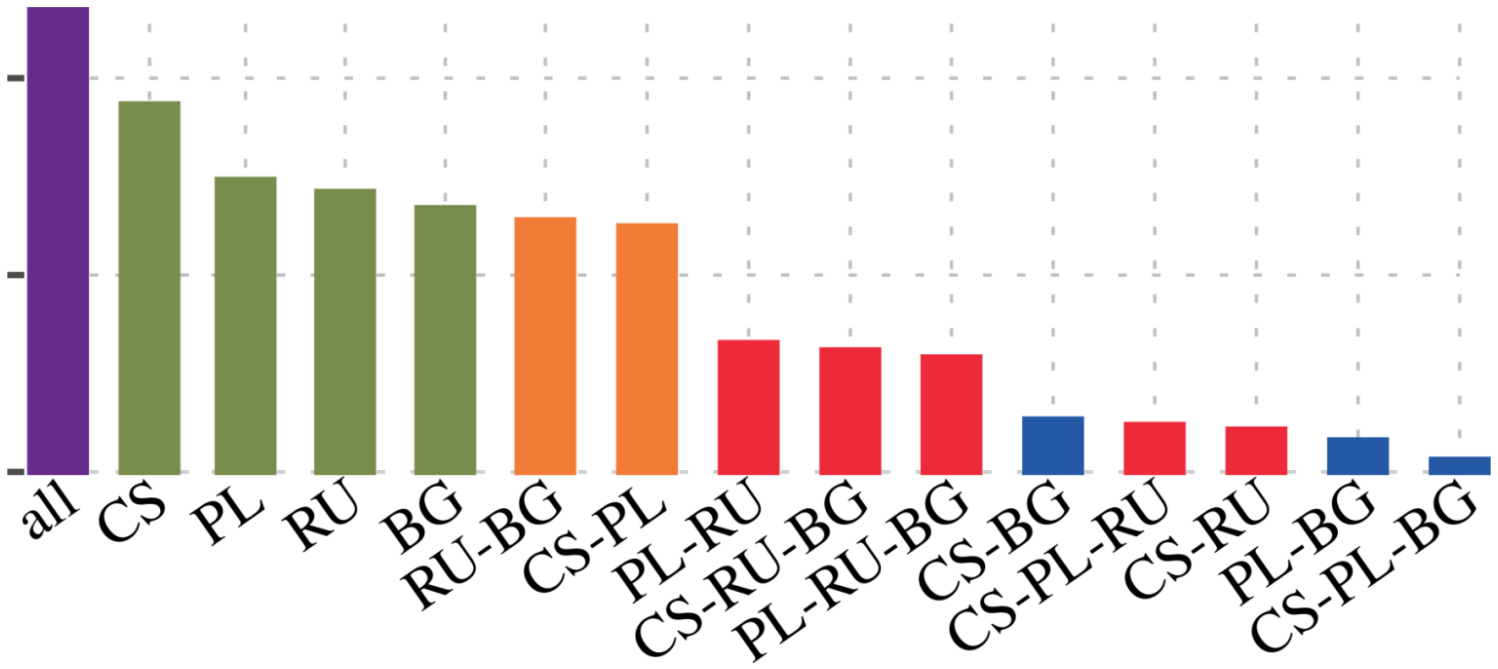
A. für phylogenetische Analyse

	usb	lsb	CS	SK	PL	SL	SC_l	SC_c	MK	BG	RU	UK	rue	BE
usb	.00	**.52**	.53	.52	.60	.57	.61	.62	*.76*	.75	.68	.70	.67	.64
lsb	**.52**	.00	.65	.66	.72	.67	.68	.71	*.87*	.85	.80	.82	.78	.74
CS	.53	.65	.00	**.41**	.56	.50	.53	.55	*.71*	.69	.61	.64	.58	.59
SK	.52	.66	**.41**	.00	.58	.48	.51	.56	*.68*	.66	.60	.65	.59	.60
PL	.60	.72	**.56**	.58	.00	.64	.64	.67	*.82*	.79	.71	.74	.69	.63
SL	.57	*.67*	.50	.48	.64	.00	**.36**	.39	.59	.58	.61	.65	.60	.61
SC_l	.61	*.68*	.53	.51	.64	.36	.00	**.04**	.54	.57	.63	.66	.62	
SC_c	.62	*.71*	.55	.56	.67	.39	**.04**	.00	.51	.53	.60	.63	.59	.59
MK	.76	*.87*	.71	.68	.82	.59	.54	**.51**	.00	.54	.74	.78	.75	.75
BG	.75	*.85*	.69	.66	.79	.58	.57	**.53**	.54	.00	.70	.77	.70	.71
RU	.68	*.80*	.61	.60	.71	.61	.63	.60	.74	.70	.00	.52	.53	**.51**
UK	.70	*.82*	.64	.65	.74	.65	.66	.63	.78	.77	.52	.00	.45	**.45**
rue	.67	*.78*	.58	.59	.69	.60	.62	.59	.75	.70	.53	**.45**	.00	.54
BE	.64	.74	.59	.60	.63	.61	.63	.59	*.75*	.71	.51	**.45**	.54	.00



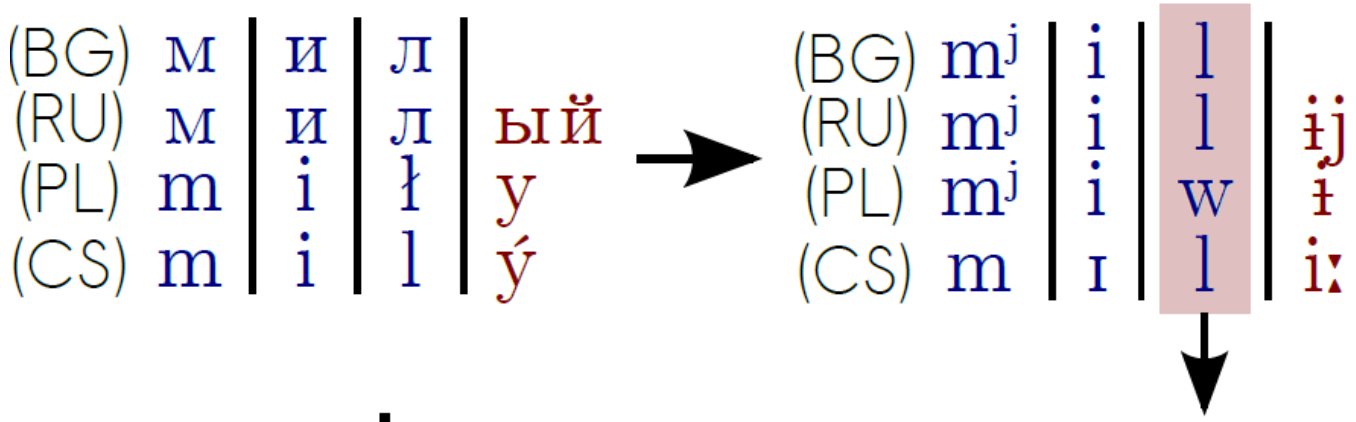
Quantifizierung sprachlicher Ähnlichkeit

B. .innerhalb von Teilmengen von Sprachen oder Sprachvarietäten



Quantifizierung sprachlicher Ähnlichkeit

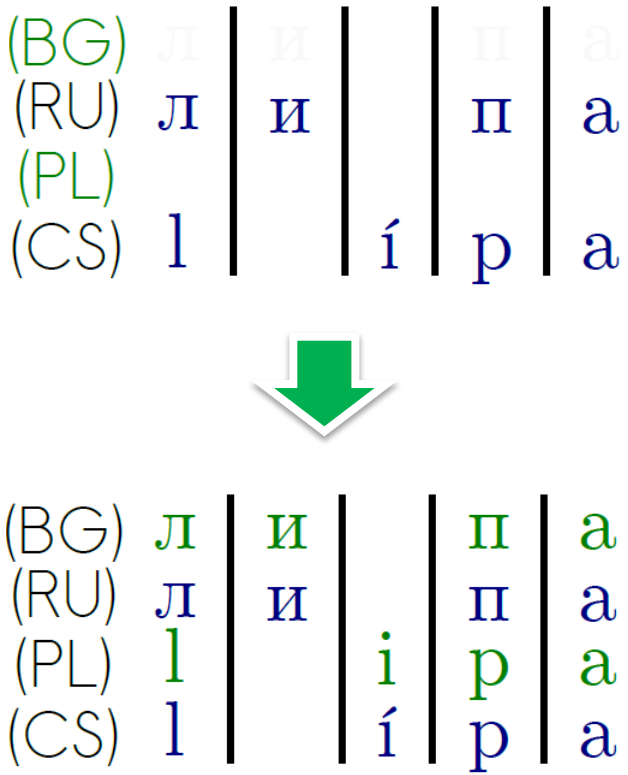
C. von Lautkorrespondenzen sowie Lautveränderungen

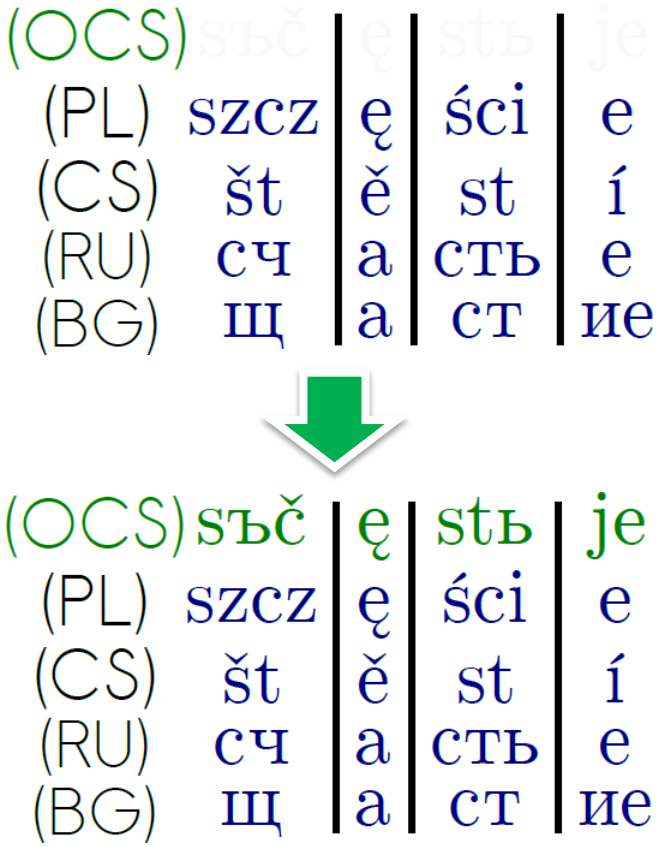


approximant, voiced, oral, central, pulmonic, **alveolar**	
approximant, voiced, oral, central, pulmonic, **alveolar**	
approximant, voiced, oral, central, pulmonic	**labial. velar**
approximant, voiced, oral, central, pulmonic, **alveolar**	

Entsprechungen finden (und verwenden)

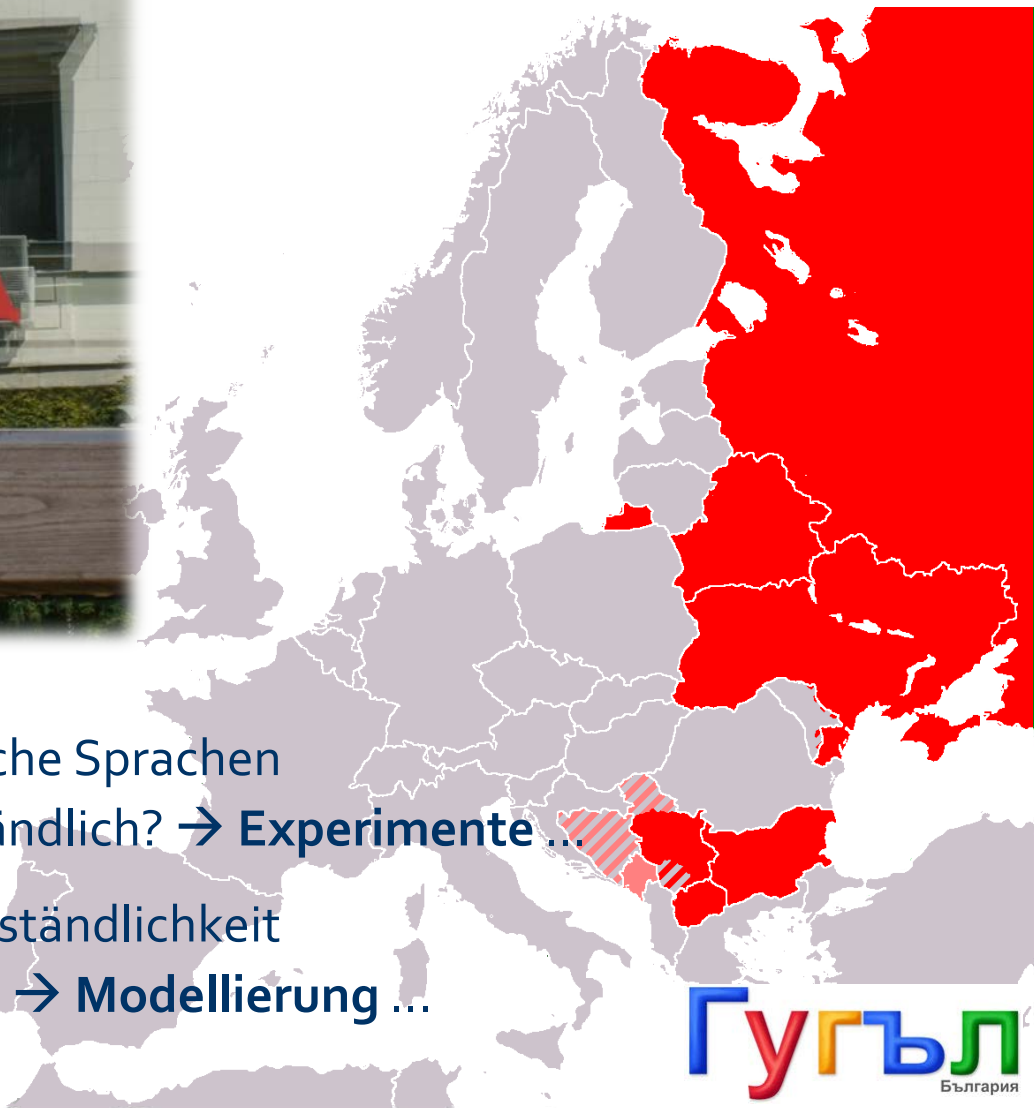
D. Reconstruktion unbekannter Formen





E. Abweichungen von der üblichen Rechtschreibung analysieren

Kyrilische Schrift



- Inwieweit sind ost- und südslawische Sprachen orthographisch gegenseitig verständlich? → **Experimente** …
- Wie kann die orthographische Verständlichkeit vorhergesagt und erklärt werden? → **Modellierung** …

Orthographische Kodierung

INCOMSLAV
SFB 1102

RU	а б в г д е ё ж з и й к л м н о п р с т у ф х ц ч ш щ ъ ы ь э ю я
UK	а б в **г ґ** д **е є** ж з **и і ї** й к л м н о п р с т у ф х ц ч ш **щ** ь ю я
BE	а б в **г** д е ё ж з **і** й к л м н о п р с т у **ў** ф х ц ч ш ы ь э ю я
BG	а б в г д **е** ж з и й к л м н о п р с т у ф х ц ч ш **щ ъ** ь ю я
MK	а б в г д **ѓ е** ж з **ѕ** и ј к л **љ** м н **њ** о п р с т **ќ** у ф х ц ч **џ** ш
SR	а б в г д **ђ е** ж з и **ј** к л **љ** м н **њ** о п р с т **ћ** у ф х ц ч **џ** ш

- Buchstaben sind identisch und ihre Lautwerte sind (in etwa) gleich
- Buchstaben sind identisch, ihre Lautwerte sind (in Abhängigkeit von der Position) nicht gleich
- Buchstaben gehören nicht zum russischen Alphabet, ihre Lautwerte sind (in der Regel) nicht bekannt

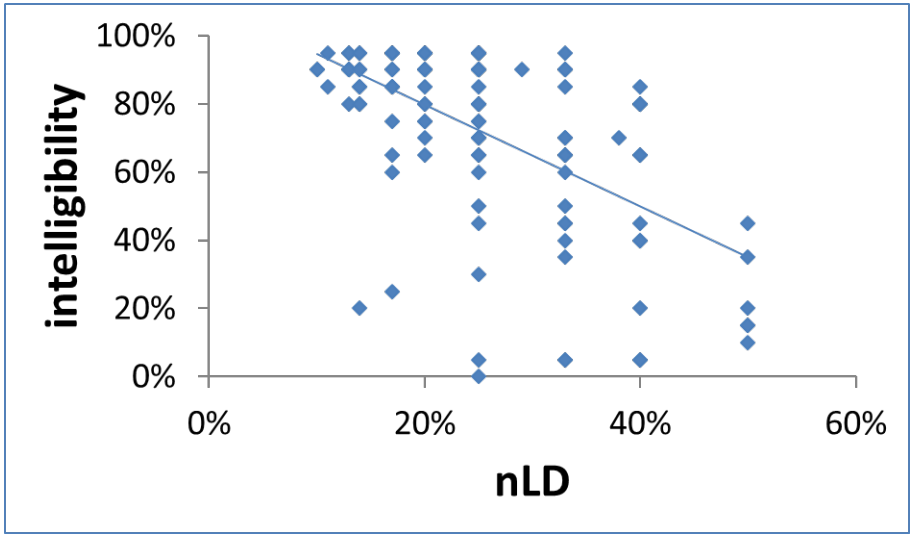
Kognaten

PS	RU	UK	BE	BG	MK	SR	EN
*synъ	сын	син	сын[1]	син	син	син	*son*
*snĕgъ	снег	сніг	снег	сняг	снег	снег	*snow*
*xlĕbъ	хлеб	хліб	хлеб	хляб	_леб	хлеб	*bread*
*melko	молоко	молоко	малако	мляко	млеко	млеко	*milk*
*berza	берёза	береза	бяроза	бреза	бреза	бреза	*birch*
*ryba	рыба	риба	рыба	риба	риба	риба	*fish*
*orьlъ	орёл	орел	арол	орел	орел	орао	*eagle*
*rǫka	рука	рука	рука	ръка	рака	рука	*hand*
*dьnь	день	день	дзень	ден	ден	дан	*day*

[1]identische Wörter wurden nicht getestet (Orientierung an das Russische)

Orthographische Distanz

- **Written word translation task (wwtt):**
  120 BG Stimuli übersetzt von
  40 RU Muttersprachlern



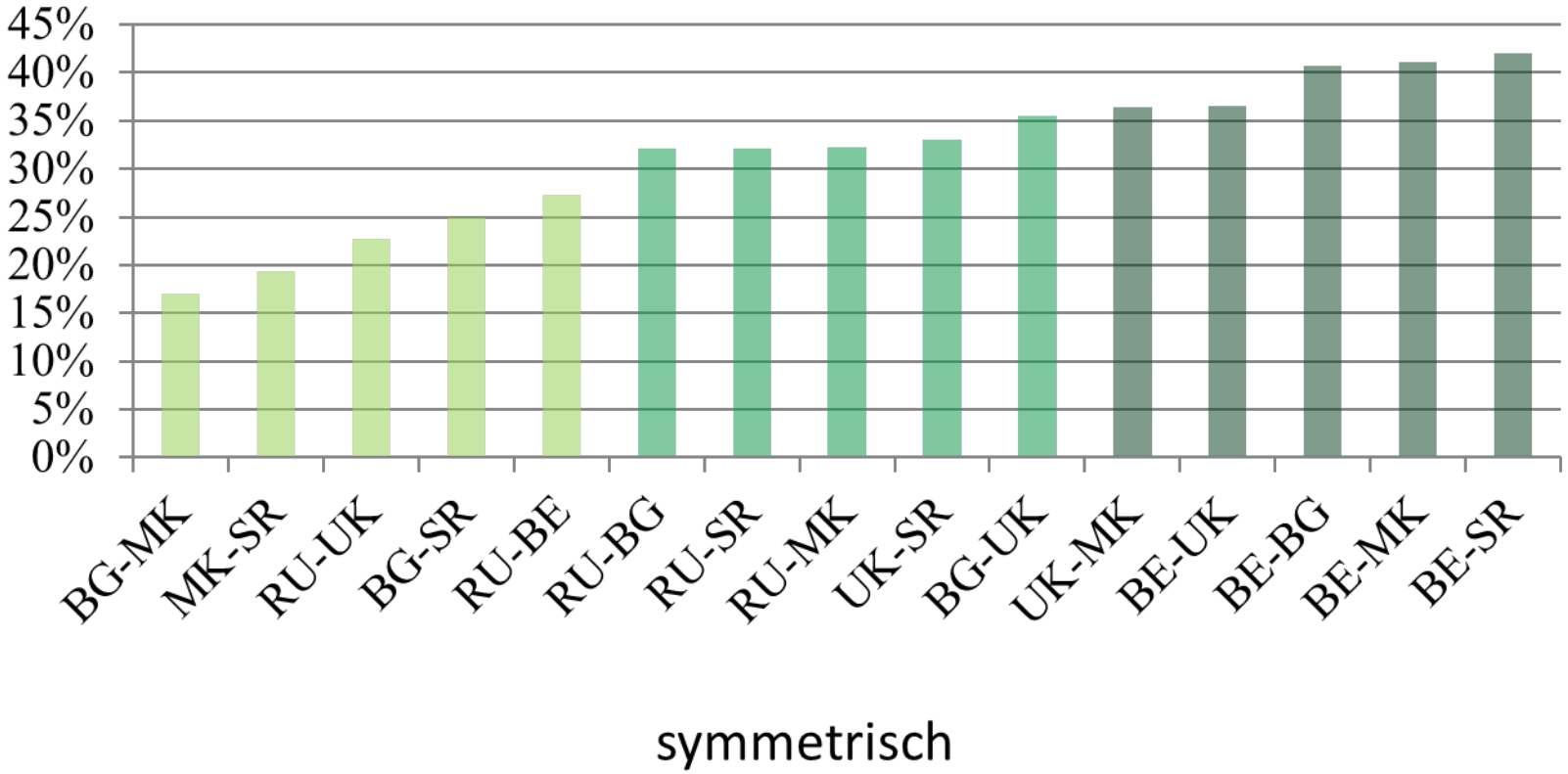
- **Levenshtein distance (LD):**
  Editierdistanz zwischn Zeichenketten
  (Einfügen, Löschen, Ersetzen)

'hunger'						normalized LD (nLD)
BG	г		л	а	д	
RU	г	о	л	о	д	
Edit costs	0	1	0	1	0	2/5→0.4

Messmethoden der orth. Verständlichkeit

- Orthographische Distanz (normierte Levenshtein-Distanz LD in %)

(Material: Carlton 1991)



Orthographische Asymmetrie

- **Character adaptation surprisal (CAS)**
  Adaptationssurprisal von Zeichen

- e.g. for Russian readers

	UK for RU	BE for RU			BG for RU			MK for RU				SR for RU					
characters of L2	a	a			a			a				a					
characters of L1	a	o	a	e	a	o	я	a	o	я	y	a	o	e	ë	я	ø
CAS values	**0**	**0.77**	**1.61**	**3.53**	**0.6**	**1.82**	**4.14**	**1.05**	**2.17**	**2.75**	**2.75**	**1.63**	**1.63**	**2.63**	**3.37**	**3.95**	**4.95**

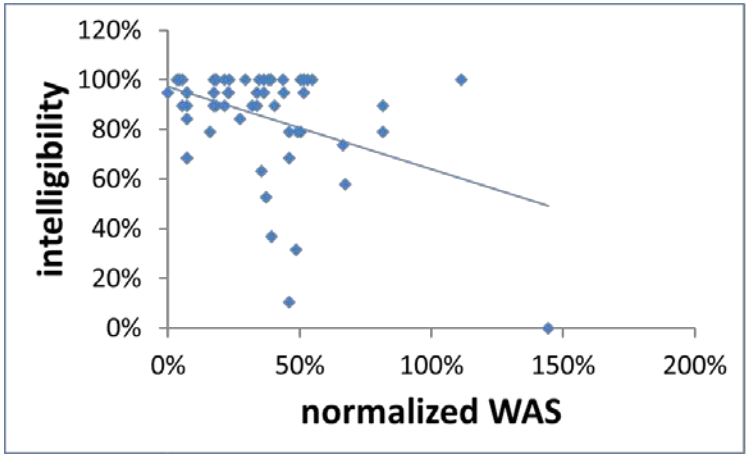
- getestet auf UK, BE, BG, MK, SR Stimuli und RU Kognaten

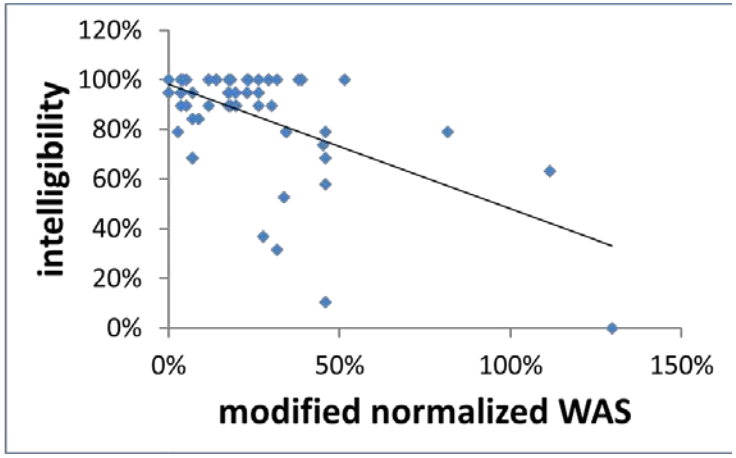
Orthographische Asymmetrie

- **Word adaptation surprisal (WAS)** → Summe von CAS im Stimuluswort

'hunger'						normalized WAS (nWAS)
BG	г		л	а	д	
RU	г	о	л	о	д	
RU reader	0	2.1	0	1.8	0	3.9/5→0.78

- **Written word translation task (wwtt):** 60 UK Stimuli übersetzt von 19 RU Muttersprachlern

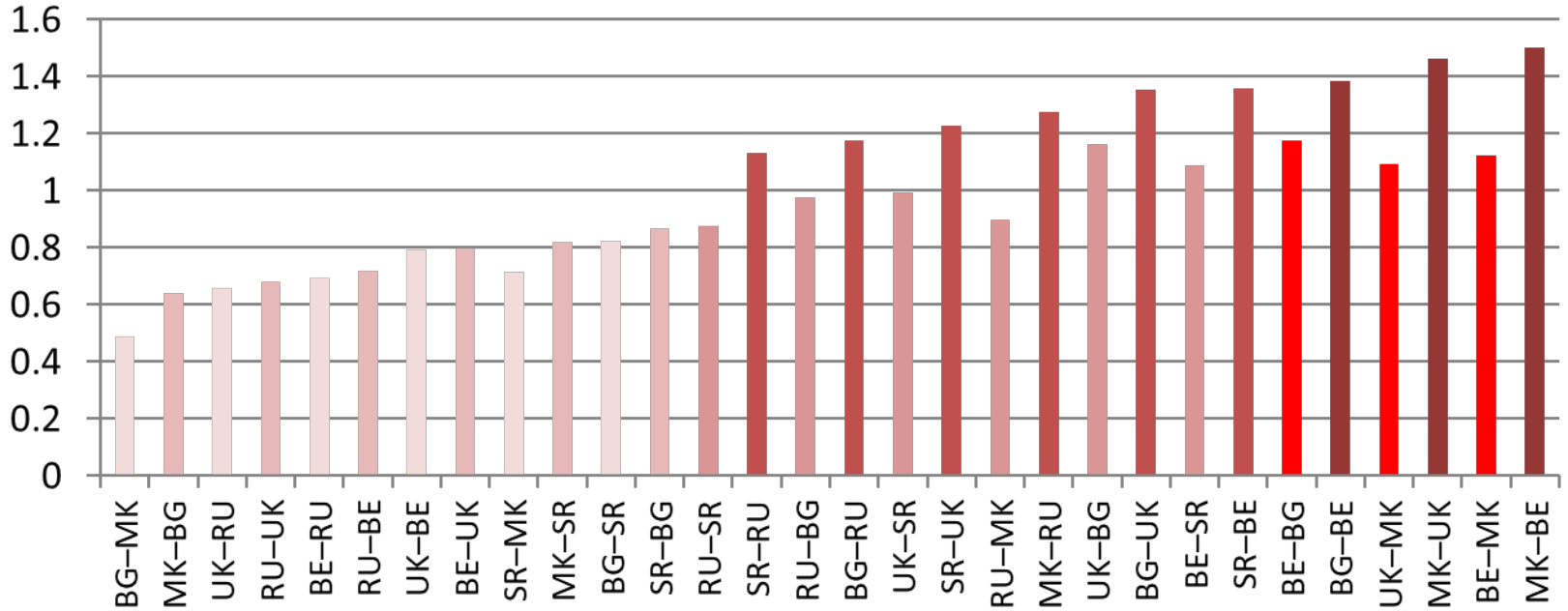




Messmethoden der orth. Verständlichkeit

- Orth. Asymmetrie (normiertes WAS in Bits)

(Material: Carlton 1991)



BG für MK < MK für BG

BE für MK < MK für BE

asymmetrisch

Empirische Untersuchung

Wie verständlich sind ost- und südslavische Sprachen für **russischsprechende** Lesende?

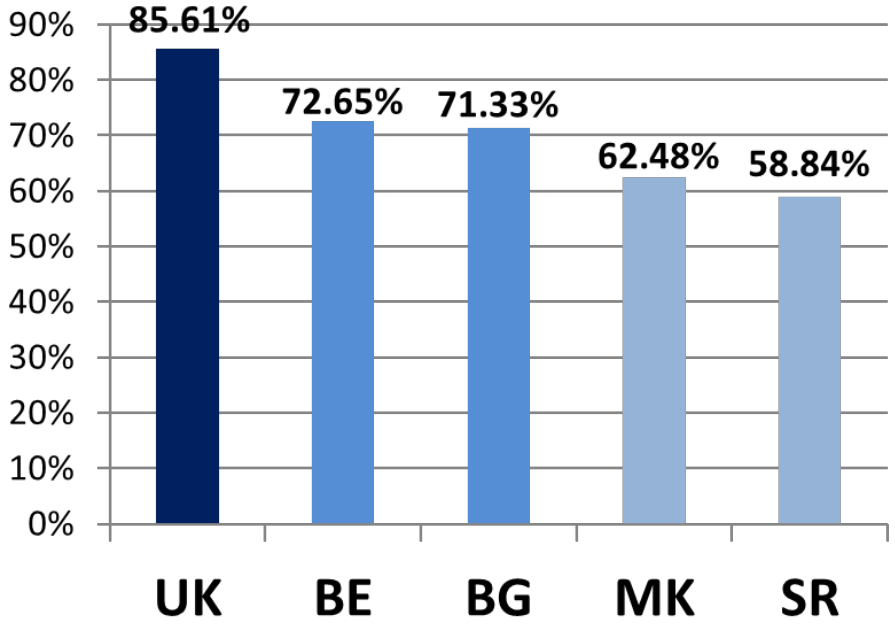
119 Probanden

¾ weiblich & ¼ männlich

Durchschnittsalter: 34 Jahre

5 Stimuli-Sprachen

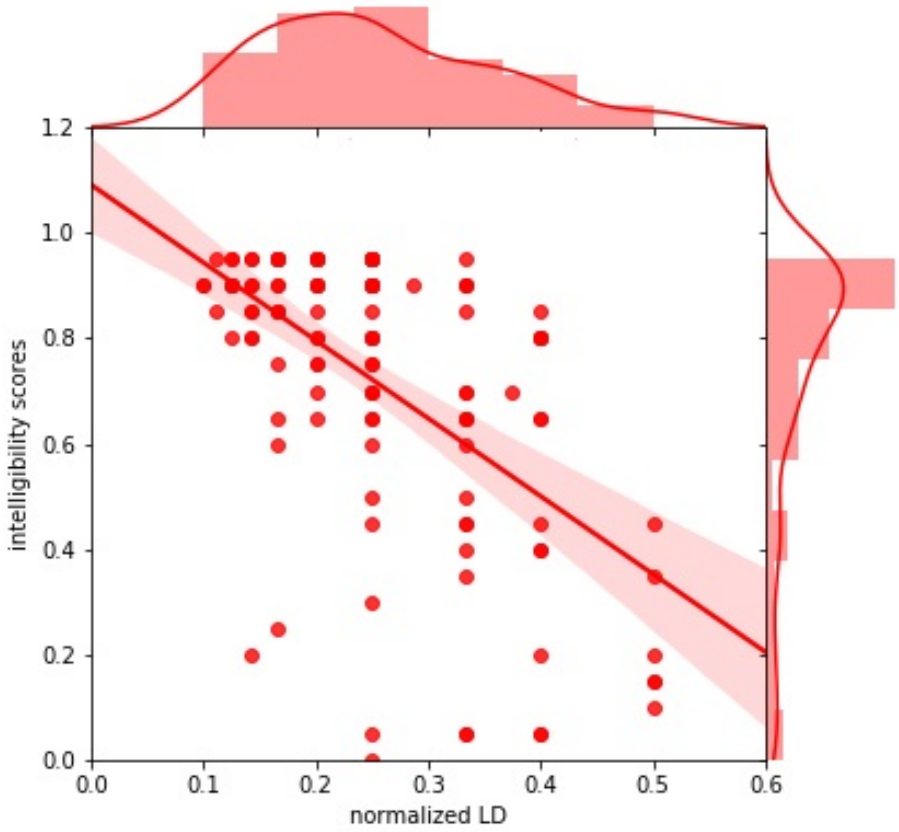
340 Stimuli-Wörter



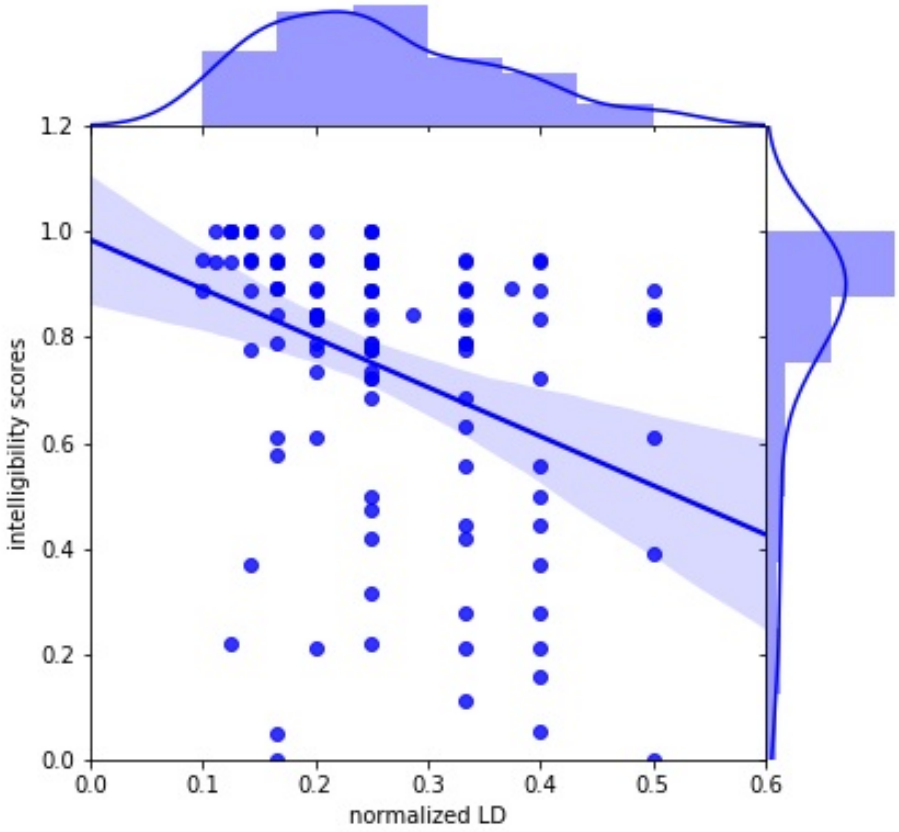
Levenshtein-Distanz (LD) als Prädiktor der gegenseitigen Verständlichkeit

INCOMSLAV
SFB 1102

- BG für RU



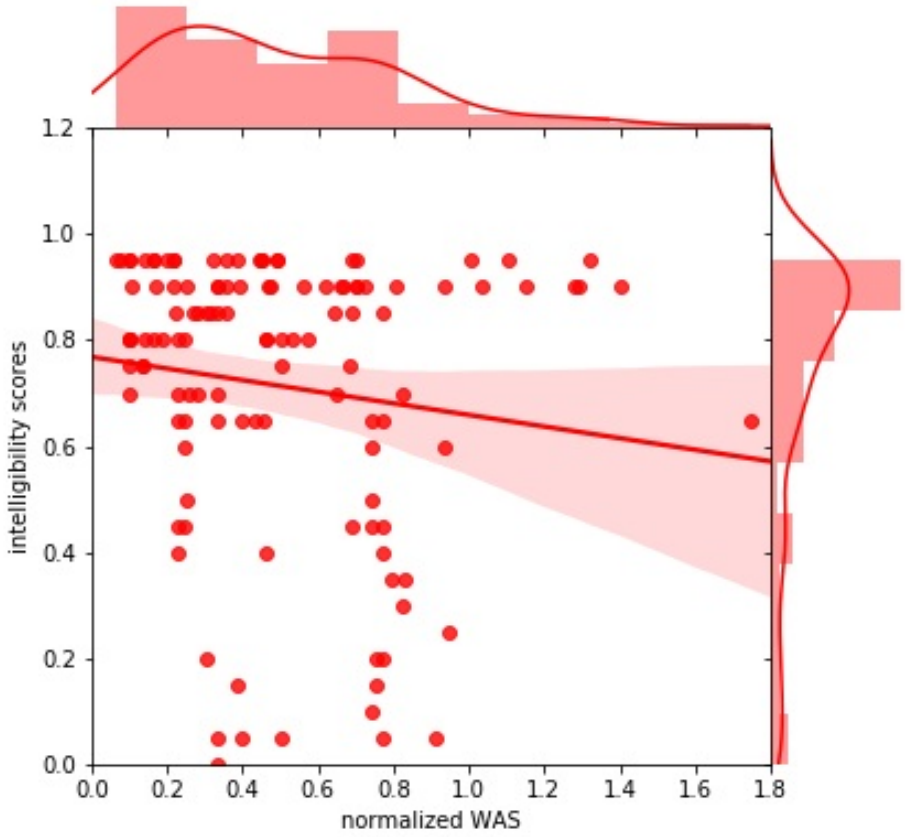
- RU für BG



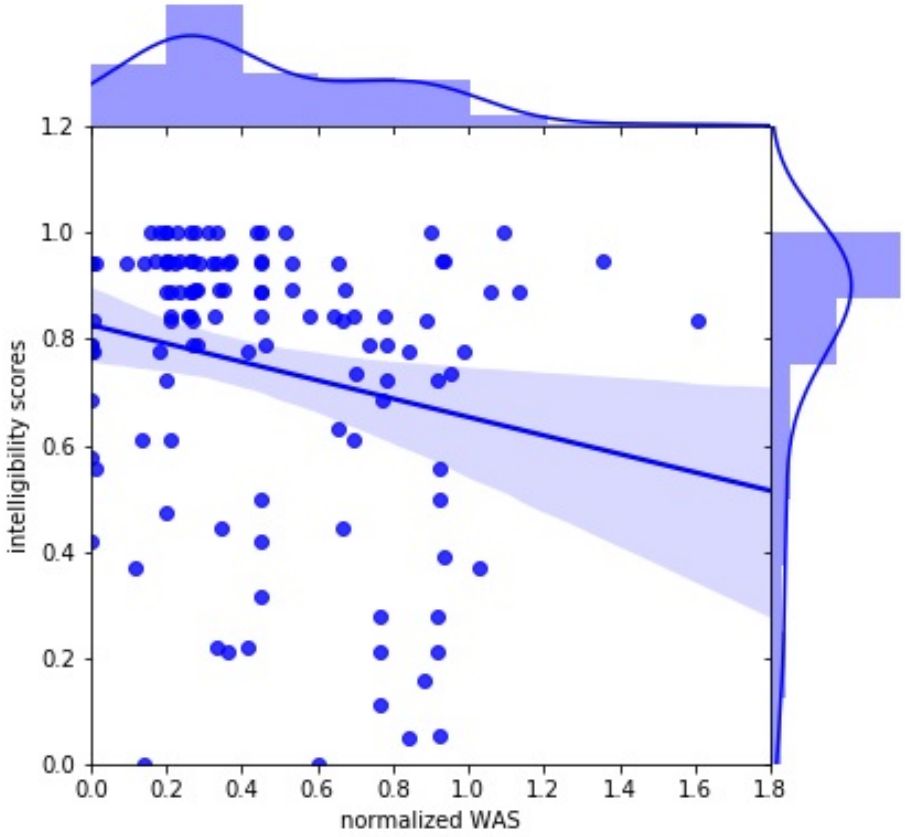
Wortadaptationssurprisal (WAS) als Prädiktor der gegenseitigen Verständlichkeit

INCOMSLAV
SFB 1102

- BG für RU



- RU für BG



Überblick

INCOMSLAV
SFB 1102

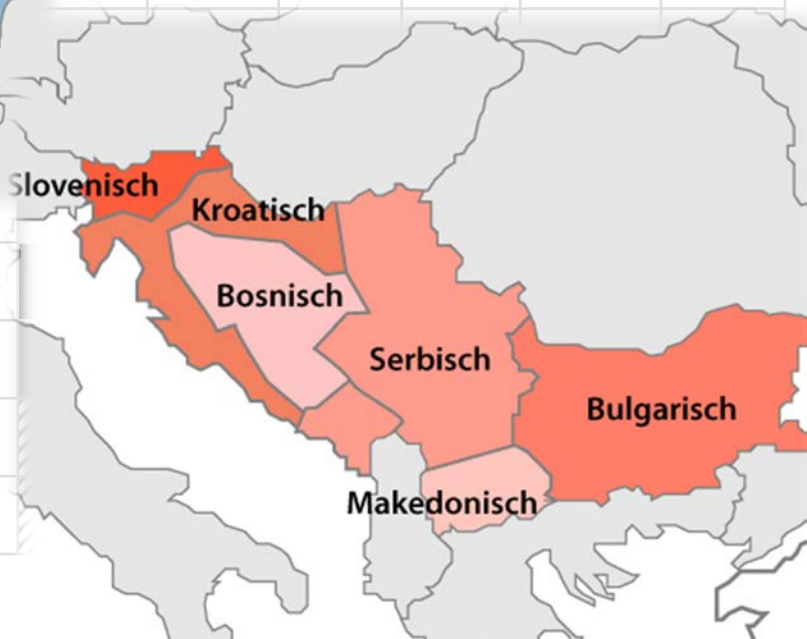
Empirische Basis		
strukturierte Daten	**„Big Data" – unstrukturiert**	**experimentelle Daten**
slavistische Expertise: historisch-vergleichend	Korpora; Wortlisten	spontane Interkomprehension

Linguistische Modellierung			
Distanzberechnungen	**statistische Verfahren**		**Informationstheorie**
Levenshtein-Metrik (Editierdistanz)	MDL	N-Gram	Entropie (Ungewissheit); Surprisal (Informationsgehalt)

Tratitionell angenommene Verständlichkeit

INCOMSLAV
SFB 1102

	:hsb	:dsb	:pol	:ces	:slk	:bos	:hrv	:srp	:slv	:mkd	:bul
pol											
ces											
slk											
bos											
hrv											
srp											
slv											
mkd											
bul											

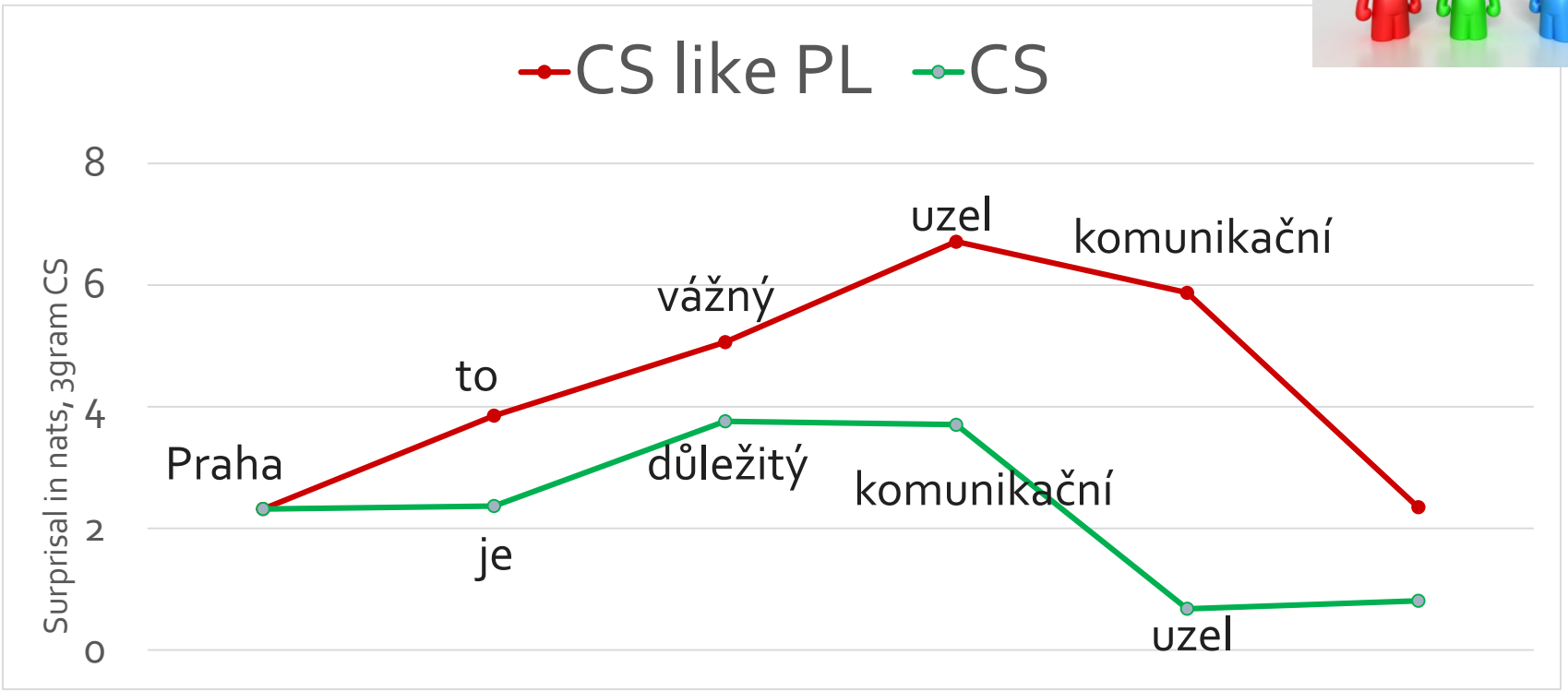


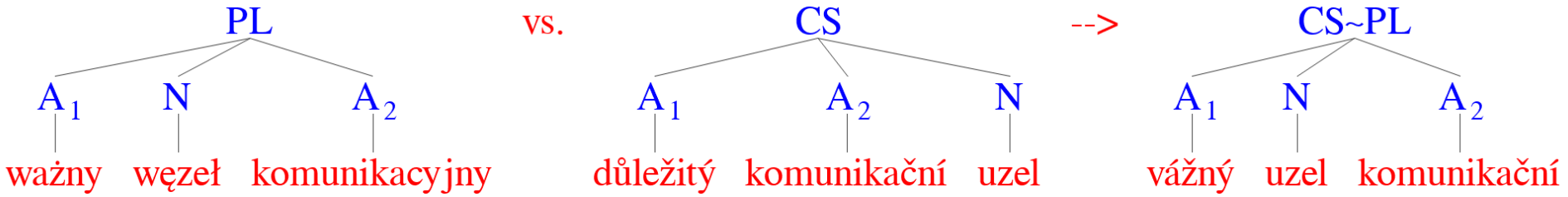
e.g. Czech LM & Cross-lingual Surprisal

INCOMSLAV
SFB 1102

LM	:rus	:ukr	:bel	:hsb	:dsb	:pol	:ces	:slk	:bos	:hrv	:srp	:slv	:mkd	:bul
rus	rus													
ukr		ukr												
bel			bel											
hsb				hsb										
dsb					dsb									
pol						pol								
ces														
slk								slk						
bos									bos					
hrv										hrv				
srp											srp			
slv												slv		
mkd													mkd	
bul														bul

PL: Praga to ważny węzeł komunikacyjny





PL-für-CS: imperfektes linguistisches Signal

- Unerwartete orthographische Einheiten in ansonsten verständlichen Wörtern: WODA statt VODA
- Unbekannte Diakritika: ą, ł, ż
- Nicht-Kognaten in ansonsten verständlichen Sätzen
- Unbekannte morphologische Einheiten
- Unerwartete Wortfolge → Unvorhersehbarkeit
- Kombinationen dieser

"The basic mission/task of the Czech-Polish/Polish-Czech Forum is to support both current and new common initiatives within the civil societies of both countries.

The Forum continues the tradition of cooperation between independent dissident groups in non-democratic times before the changeover of 1989, which culminated in the activities of the Polish-Czech-Slovak Solidarity movement." (http://www.mzv.cz/cesko-polske_forum/)

Hypothesen

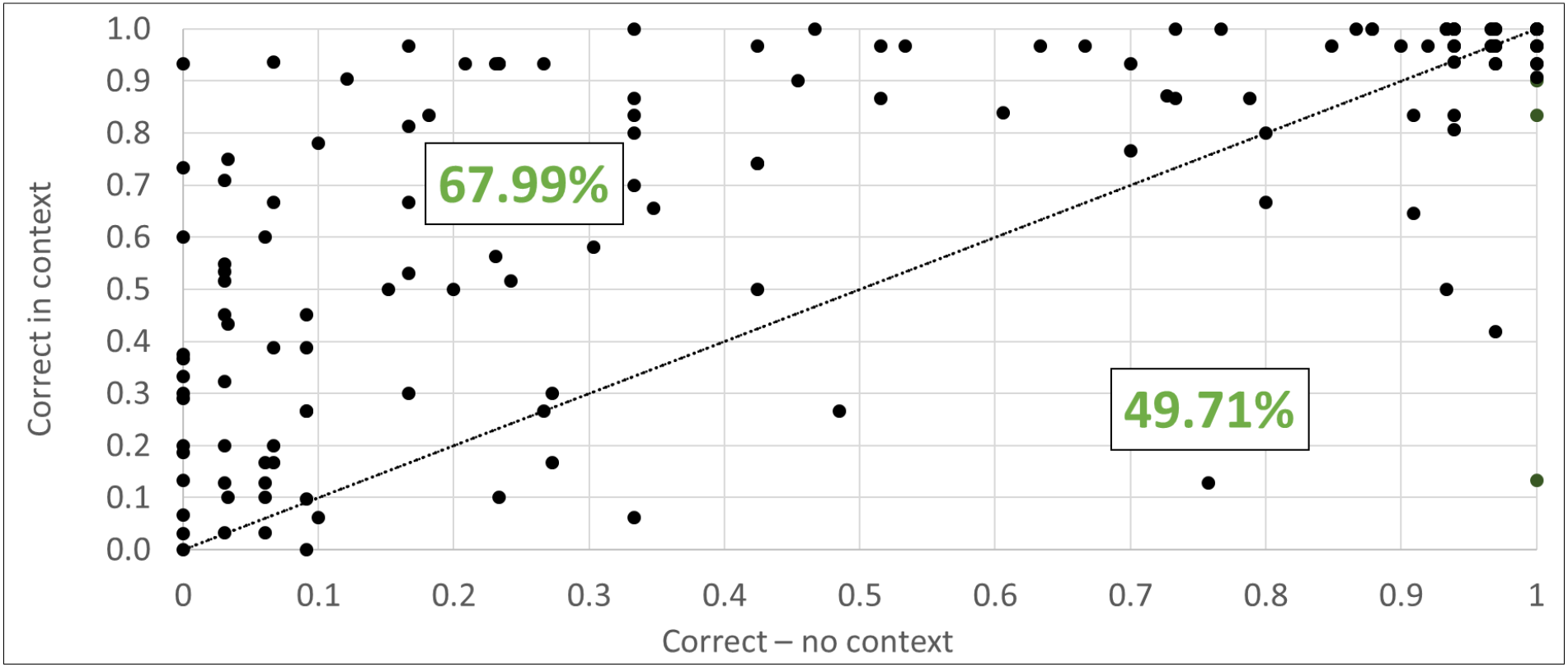
1. Verständlichkeit von Zielwörtern ist größer, wenn diese im vorhersehbaren Satzkontext präsentiert werden, als ohne Kontext.

2. Nicht nur sprachliche Distanz, sondern auch Surprisal von 3-gram Modellen korreliert mit der Verständlichkeit von Zielwörtern

Ergebnisse mit vs. ohne Kontext

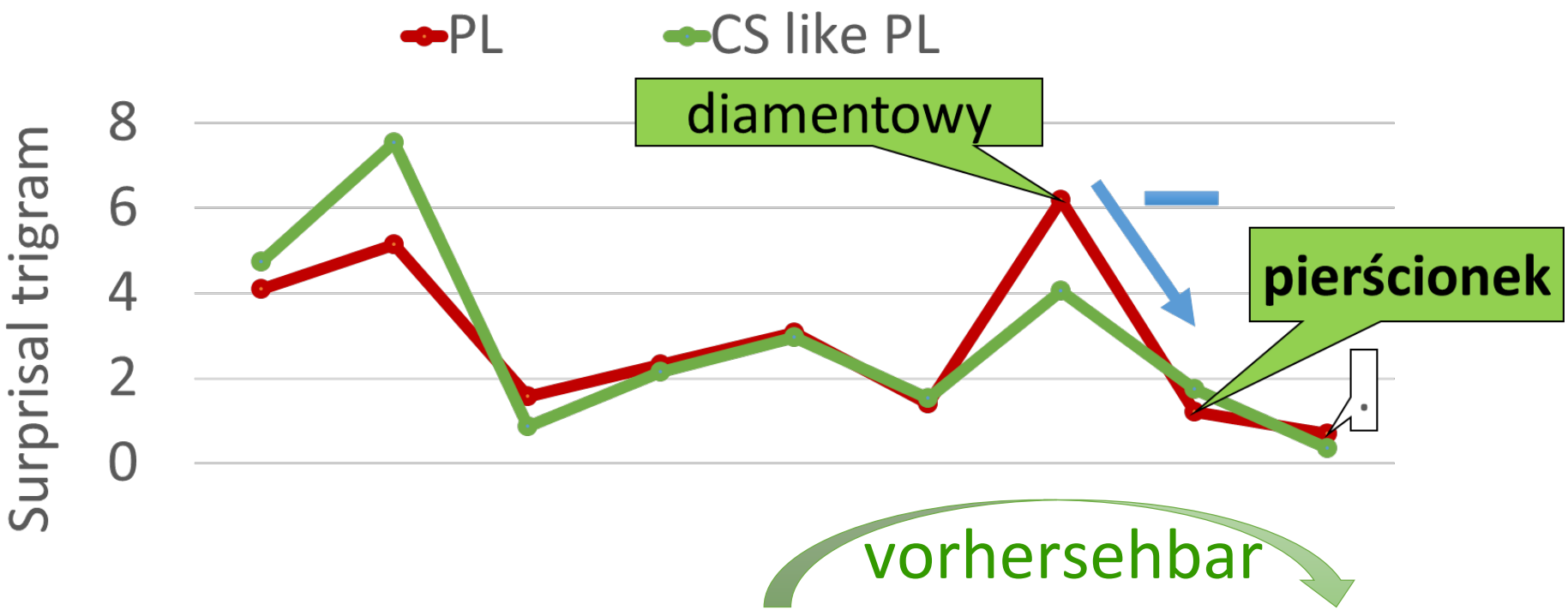
INCOMSLAV
SFB 1102

Hypothese 1 ✓



(Jágrová & Avgustinova, CICLing 2019)

Hilfreiches Wort vor Zielwort

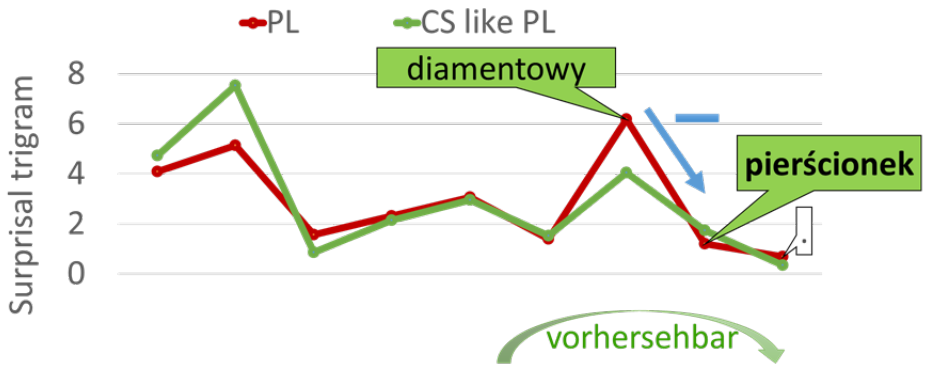


PL: *Bob oświadczył się i dał jej **diamentowy pierścionek**.*

CS: *Bob se zasnoubil a dal jí **diamantový prstýnek**.*

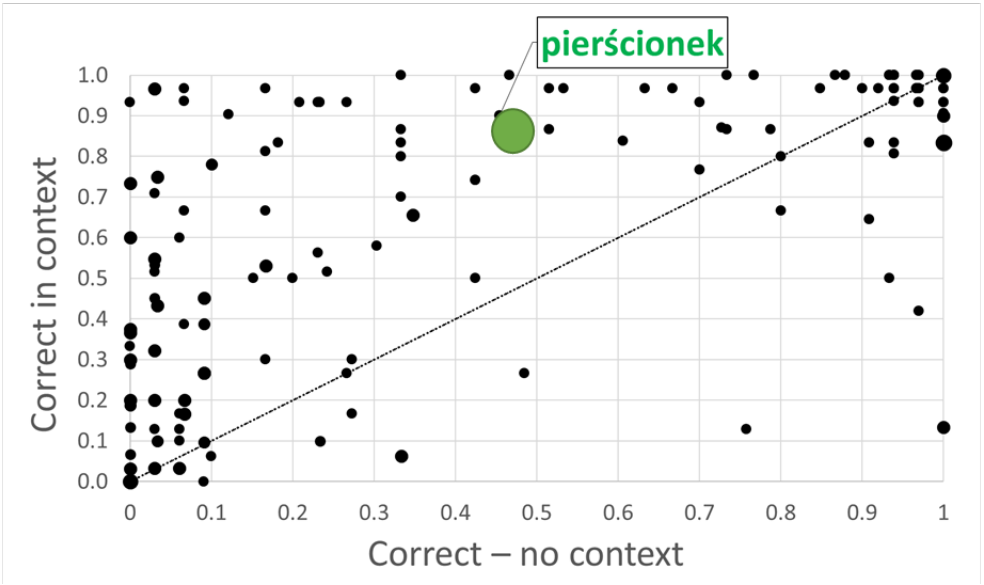
EN: *Bob proposed and gave her a **diamond ring**.*

Surprisal von 3-gram Modell



PL: *Bob oświadczył się i dał jej **diamentowy pierścionek**.*
CS: *Bob se zasnoubil a dal jí **diamantový prstýnek**.*
EN: *Bob proposed and gave her a **diamond ring**.*

Hypothese 2 ✓



Überblick

INCOMSLAV
SFB 1102

Empirische Basis		
strukturierte Daten	**„Big Data" – unstrukturiert**	**experimentelle Daten**
slavistische Expertise: historisch-vergleichend	Korpora; Wortlisten	spontane Interkomprehension

Linguistische Modellierung			
Distanzberechnungen	**statistische Verfahren**		**Informationstheorie**
Levenshtein-Metrik (Editierdistanz)	MDL	N-Gram	Entropie (Ungewissheit); Surprisal (Informationsgehalt)

Zusammenfassung

- **Rezeptive Mehrsprachigkeit** wird ermöglicht durch die menschliche Fähigkeit, **imperfekte Sprachsignale** robust zu verarbeiten.

- **Verarbeitungsaufwand** bei **Interkomprehension** kann in zwei Dimensionen erfasst werden:

  1. **linguistische Distanz**

     als symmetrischer Maß von sprachlicher Ähnlichkeit

  2. **Surprisal** (bzw. Vorhersgbarkeit im Kontext)

     als asymmetrischer Maß von Informationsdichte

Slavische Interkomprehensionsmatrix Online

http://intercomprehension.coli.uni-saarland.de/en/SlavMatrix/Results/

INCOMSLAV SFB 1102

Informationstheoretische Modellierung: Ähnlichkeit + Asymmetrie

Levenshtein-Distanz

	1	2	3	4	LD
BG	е	з	и	к	
RU	я	з	ы	к	
BG für RU	1	0	1	0	2
RU für BG	1	0	1	0	2

Ø nLD von 120 BG-RU Kognaten: 25,61%

Bedingte Entropie

	1	2	3	4	BE
BG	е	з	и	к	
RU	я	з	ы	к	
BG für RU	1,65	0	0,96	0	2,61
RU für BG	1,42	0	0	0	1,42

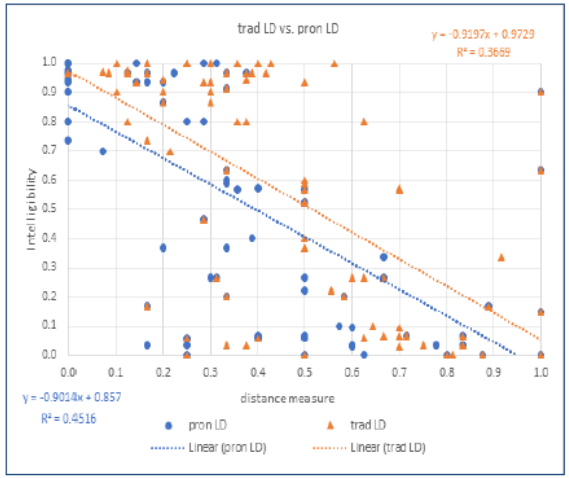
0,49 BG für RU > 0,47 RU für BG

Wortadaptationssurprisal

	1	2	3	4	WAS
BG	е	з	и	к	
RU	я	з	ы	к	
BG für RU	2,5	0	0,7	0	3,2
RU für BG	0,9	0	0	0	0,9

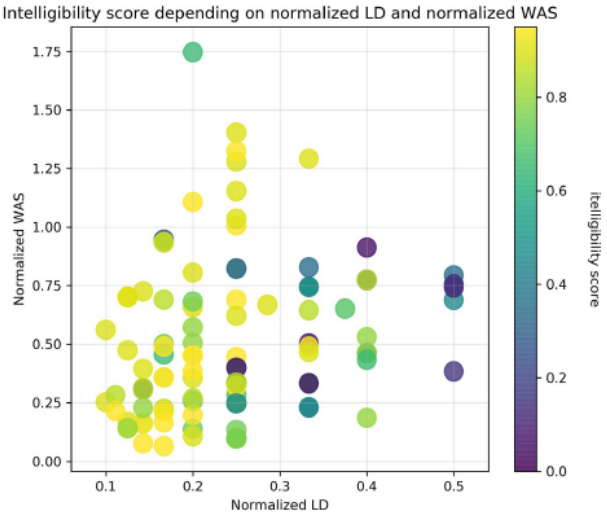
0,50 BG für RU > 0,46 RU für BG

Empirische Validierung: Korrelation mit experimentellen Daten

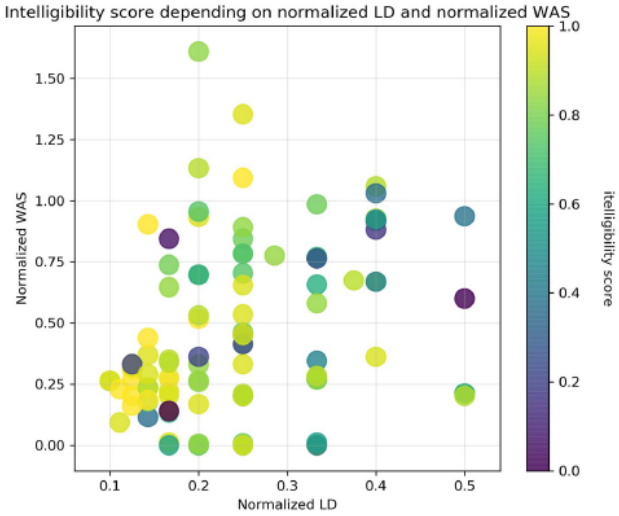


84/100 häufigste Substantive des PL

PL für CS Lesende



BG für RU Lesende



RU für BG Lesende

Asymmetrische Verständlichkeit

nLD + nWAS

BG für RU: $R^2 = 0.32$

RU für BG: $R^2 = 0.14$