Slavic Diachronic Corpora: Challenges and Perspectives

Project INCOMSLAV
Mutual Intelligibility and Surprisal in Slavic Intercomprehension

Historical Corpus Linguistics: Methods and Applications

Saarbrücken, 16-17 June 2016

Research Group

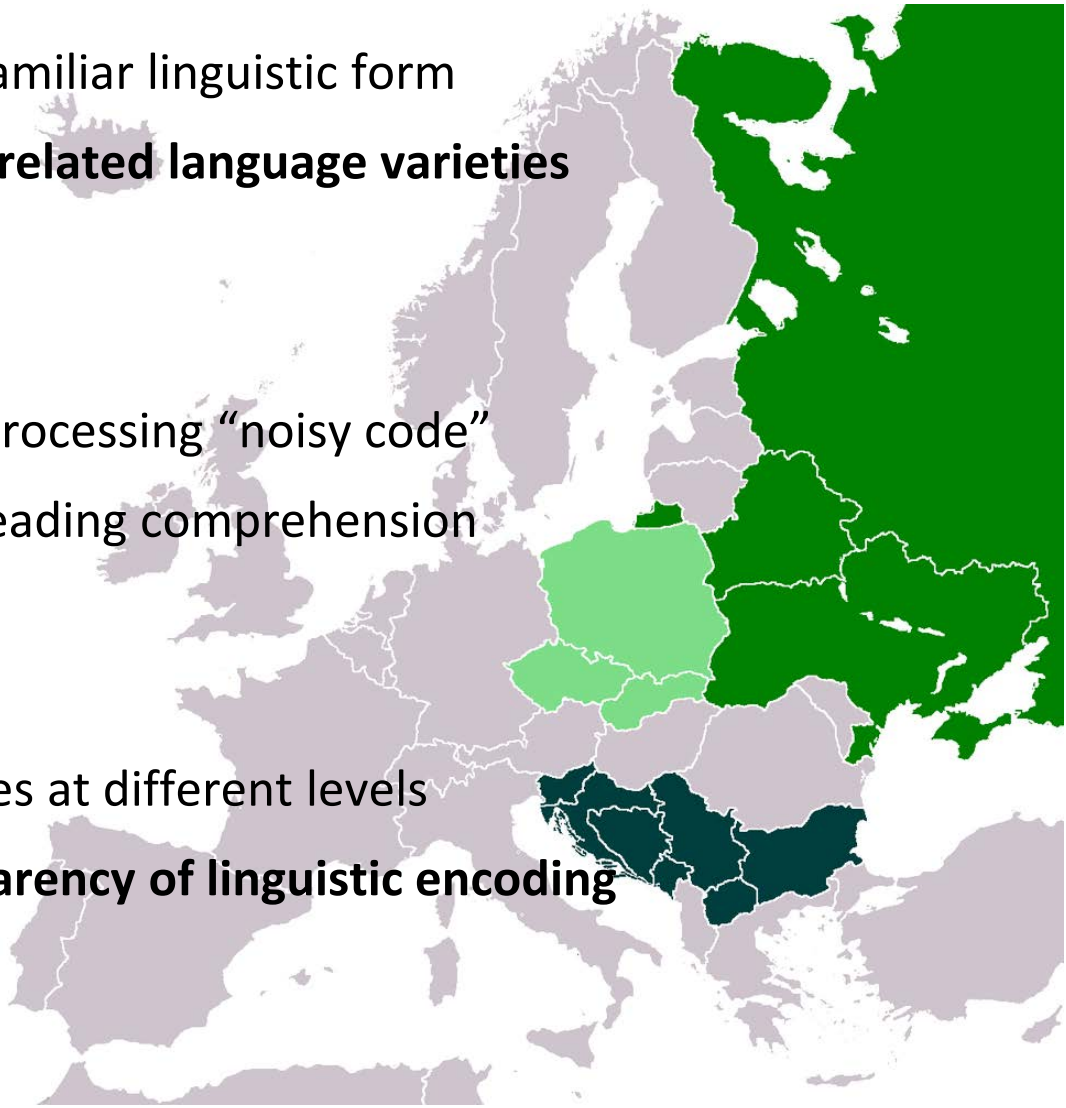
Statistical NLP

Slavonic Studies

Computational & Slavic Linguistics

Focus on Slavic Intercomprehension

- Receptive multilingualism
  - inter-lingual tolerance to unfamiliar linguistic form
  - ability to understand texts in **related language varieties**
- Surprisal
  - information-theoretic view: processing “noisy code”
  - **written input**: cross-lingual reading comprehension
- Mutual intelligibility
  - measurable linguistic distances at different levels
  - basic factor to model: **transparency of linguistic encoding**



Slavic Intercomprehension Matrix

related language variertes
written input
transparency of linguistic encoding

	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. Belorusian	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)	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(6)	9(7)	10(7)	bos	9(10)	9(11)	9(12)	9(13)	8(14)
10. Croatian	10(1)	10(2)				(6)	10(7)	11(7)	10(9)	hrv	10(11)	10(12)	10(13)	9(14)
11. Serbian	11(1)	1				6)	11(7)	11(8)	11(9)	11(10)	srp	11(12)	11(12)	10(14)
12. Slovene	12(1)	12				2(6)	12(7)	12(8)	12(9)	12(10)	12(11)	slv	12(13)	11(14)
13. Macedonian	13(1)					13(6)	13(7)	13(8)	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

How can a Bulgarian understand Russian?

Notation: A(B)

A = *decoder's language*; B = *language of the stimulus*

The diachronic dimension

related language variertes
written input
transparency of linguistic encoding

- Language-internal (direct): languages change in time
- Cross-linguistic (indirect): in relation to a common ancestor

		Church Slavonic			
Proto-Slavic 6 BC – 6 AD	East	Old Russian (X-XV)	Middle Russian (XV-XVII)	Modern Russian	Cyrillic script
	South	Old Bulgarian / OCS (IX-XI)	Middle Bulgarian (XII-XVIII)	Modern Bulgarian	
	West	Old Polish (XII-XV)	Middle Polish (XVI-XVIII)	Modern Polish	Latin script
		Old Czech (X-XV)	Middle Czech (XVI-XVIII)	Modern Czech	

From Proto-Slavic to Modern Slavic

related language variertes
written input
transparency of linguistic encoding

Latin script		← →	Cyrillic script			
PL	CZ	Proto-Slavic	OCS	RU	BG	
brat	bratr	**brat(r)ъ*	братъ	брат	брат	*brother*
syn	syn	**synъ*	сынъ	сын	син	*son*
dom	dům	**domъ*	домъ	дом	дом	*house*
rzeka	řeka	**rěka*	рѣка	река	река	*river*
śnieg	sníh	**sněgъ*	снѣгъ	снег	сняг	*snow*
chleb	chléb	**xlěbъ*	хлѣбъ	хлеб	хляб	*bread*
wino	víno	**vino*	вино	вино	вино	*wine*
woda	voda	**voda*	вода	вода	вода	*water*
ryba	ryba	**ryba*	рыба	рыба	риба	*fish*
oko	oko	**oko*	око	око	око	*eye*
ręka	ruka	**rǫka*	рѧка	рука	ръка	*hand*
żyć	žíti	**žiti*	жити	жить	живея	*live*
biały	bílý	**bělъ(jъ)*	бѣлъ	белый	бял	*white*

Diachronic and synchronic variants

related language variertes
written input
transparency of linguistic encoding

- e.g. middle PL: więtszy modern CZ: větší (bigger)
  modern PL: większy

  → middle PL closer to modern CZ

- transformable by diachronically-based cross-lingual correspondence rules
- will be tested in experiments with native speakers

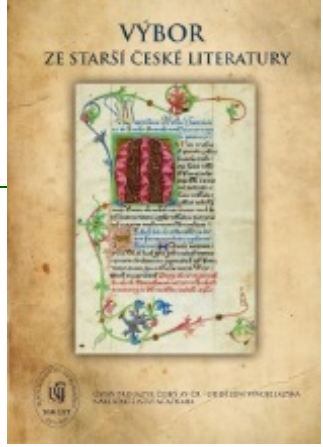
Orthography as primary interface

related language variertes
written input
transparency of linguistic encoding

- **Orthographic correlates** (used in linguistic analyses of inter-lingual similarity)
  - **in Slavic vocabulary** (common heritage): historical correspondence rules
  - **in internationalisms** (modern vocabulary): diff. in modern orthographies
  - **in morphology**: inflectional and derivational

- Major spelling issues in historical corpus linguistics
  - **Difference**: historical spelling differs from modern spelling (diachronic)
  - **Variance**: historical spelling is variable and inconsistent (synchronic)
  - **Uncertainty**: digital text is result of interpretation and transcription, which introduces artefacts and errors

Slavic diachronic corpora

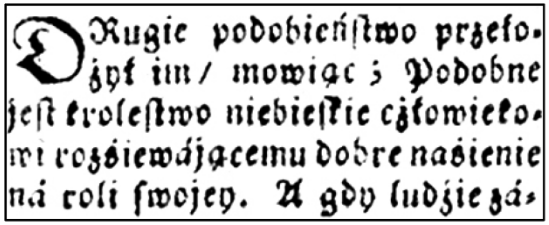


- DIAKORP (CZ) https://ucnk.ff.cuni.cz/english/diakorp.php
- Vokabulář webový (CZ)

...

- PolDi (PL) http://rhssl1.uni-regensburg.de/SlavKo/korpus/poldi
- Korpus tekstów staropolskich do roku 1500 (PL)

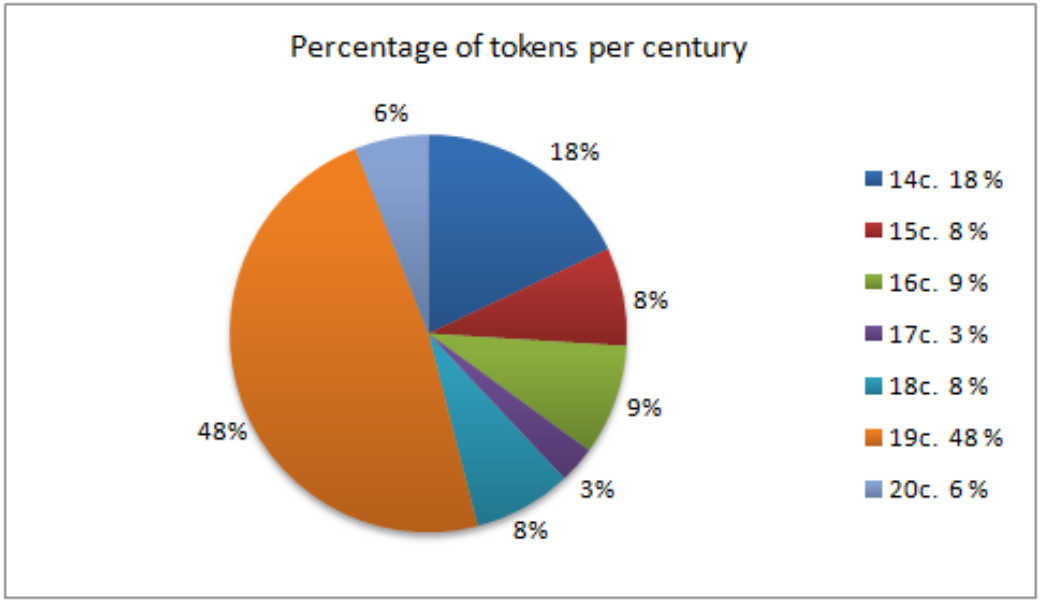
...


Drugie podobieństwo przełożył im, mówiąc: Podobne jest królestwo niebieskie człowiekowi, rozsiewającemu dobre nasienie na roli swojej.

- RRuDi (RU) http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new
- RNC: Diachronic corpus (RU)
  - Old Russian & **Birch bark letters**
  - Church-Slavonic
  - Middle Russian

e.g. Diachronic section of the Czech National Corpus

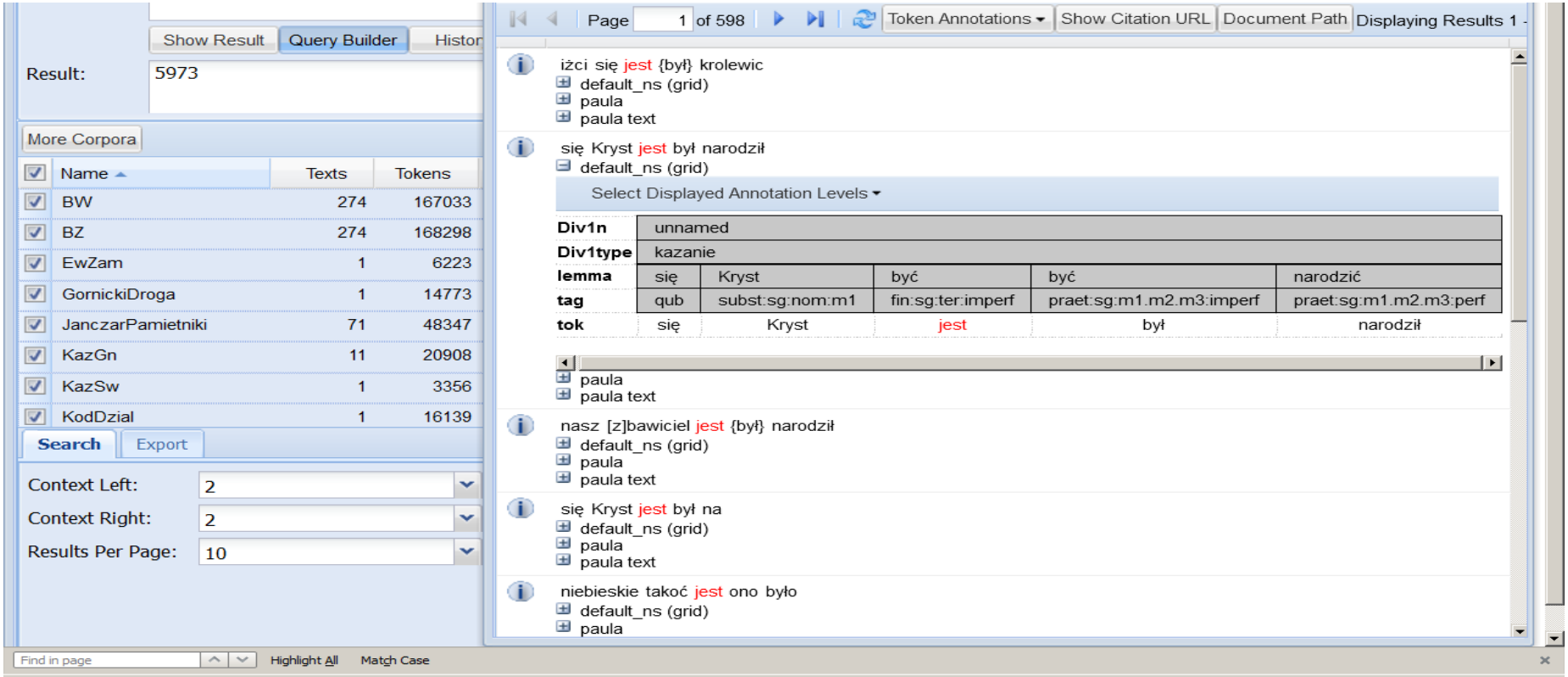
- http://wiki.korpus.cz/doku.php/en:cnk:diakorp



- different spelling systems: **simple**, **digraphic**, **diacritical** & combinations thereof
- **transcribed, not transliterated**: enabling search as in the synchronic sections
- **tagged**: to preserve certain information, which is lost when transcribing
- **hyperlemmata** to allow variety-independent search, e.g. use hyperlemma *kůň* to also find older Czech forms *kóň* and *kuoň*

e.g. Polish Diachronic Online Corpus

- tools for modern Polish + manual annotation
- **Morfeusz** as external “generic tagger" patched up with post-processing rules
- **Annis-2** as database and web interface – to visualize and make queryable “complex multilevel linguistic corpora with diverse types of annotation”



e.g. Old Russian section of the Russian National Corpus

Древнерусский корпус

Орфография: точная упрощенная модернизированная

Поиск точных форм ? А Б В

Слово или фраза

искать очистить

Русско-церковнославянская клавиатура

Упрощенная орфография

а б в г д е,є ж з ѕ и і,ї ѵ,ѷ к л м н о,ѻ ѡ,ѿ ѽ п р
с т у,оу ф ѳ х ц ч ш щ ъ ы ь ѣ ю я,ѧ ѯ ѱ ҁ

1. Александрия [омонимия снята] Все примеры (2)

, аки молнна. въѯалкавъ же сѧ, въсхотѣ(χ) пр҃ѣ҃ти хлѣба, и приѯвахъ [Александрия] [омонимия снята]
боудетъ. се же рекъ, александръ принесе къ даньдамью злато, и хлѣбы. [Александрия] [омонимия снята]

2. Волынская летопись [омонимия снята] Все примеры (1)

лноу. а по стоу. хлѣба. а по пѧти цебровъ [Волынская летопись] [омони

3. Изборник [омонимия снята] Все примеры (12)

въ свою обитѣль · обыщь ти боуди съ нимь хлѣбъ твои · обыштѧ чашѧ [Изборник] [омонимия снята]
трапезою · помяни соухъ хлѣбъ гадоуштааго · и не могоуштааго [Изборник] [омонимия снята]

хлѣбы	
Лемма	хлѣбъ
Грамматика	сущ, м, мн, вин
Доп. признаки	086:20, bcomma, bmark, last

Сообщить об ошибке...

Overview of project activities

- Establishing **orthographic correlates**
  - Czech↔ Polish; Bulgarian↔ Russian
  - **informed by comparative historical linguistic studies**
- Collecting and preparing **parallel lexical recourses**
  - Pan-Slavic vocabulary; internationalisms; Swadesh lists
  - 100 most frequent nouns extracted from national corpora (CZ, PL, RU, BG)
- Computational **transformation experiments**
  - applying diachronically-based orthographic correspondence rules on parallel word sets
  - obtaining additional statistical orthographic and morphological correspondences via MDL model

Diachronically motivated regular correspondences

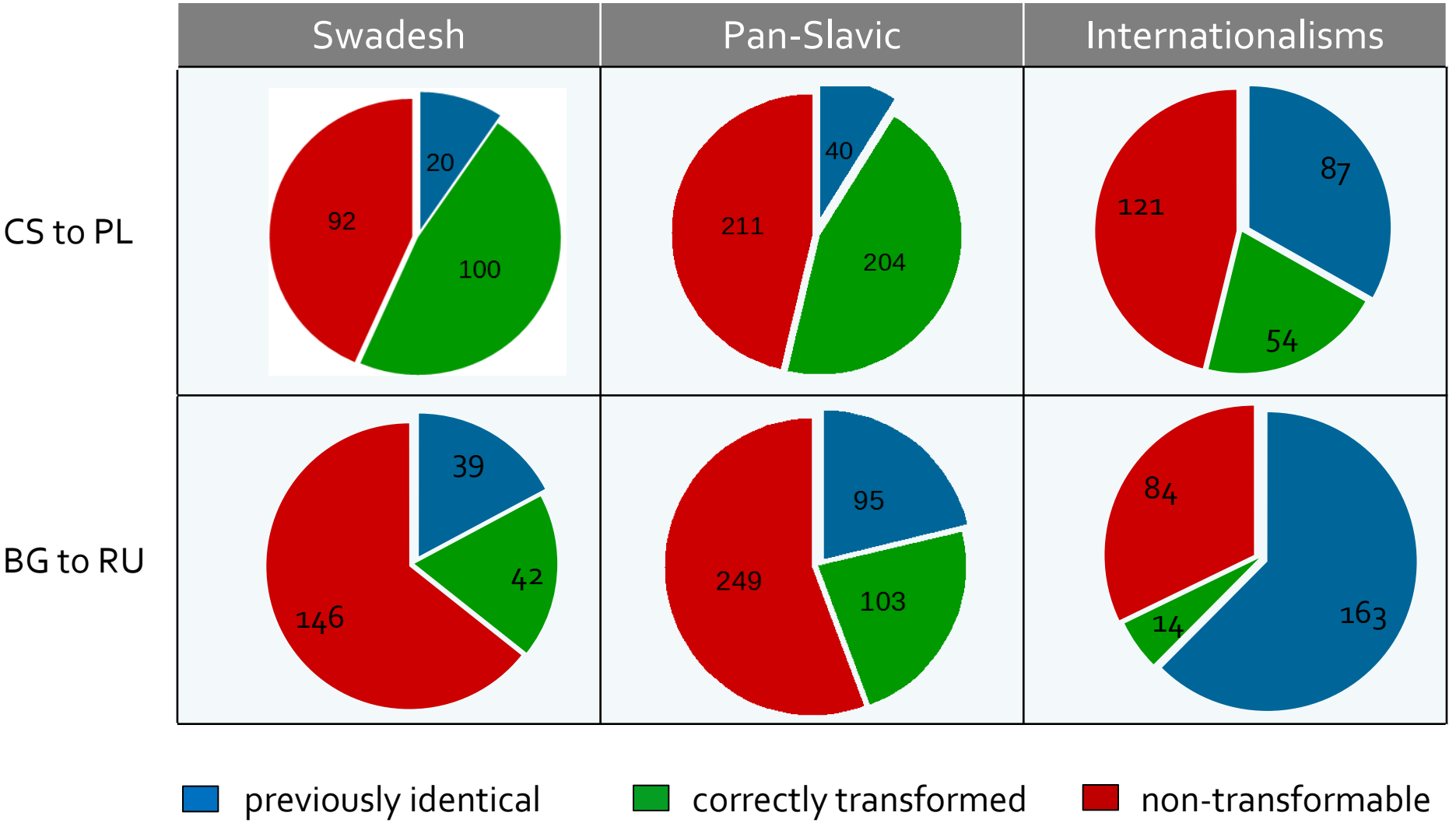
	Czech	Polish	Bulgarian	Russian
horse	kůň	koń	кон	конь
body	tělo	ciało	тяло	тело
sea	moře	morze	море	море
brush	štětka	szczotka	четка	щётка
cow	kráva	krowa	крава	корова
before	před	przed	пред	перед
head	hlava	głowa	глава	голова
voice	hlas	głos	глас	голос
full	plný	pełny	пълен	полный
yellow	žlutý	żółty	жълт	жёлтый
wolf	vlk	wilk	вълк	волк

la ło ла оло

l eł ъл ол

l il ъл ол

Results of applying linguistic rules on parallel word sets



Methodological considerations

- Diachronic linguistics aligns cognate words, looking for **regular segmental correspondence** (in order to identify sound equivalences)
  - Can the recognition of semantically related words be improved?
  - Can alignment be made more sensitive to phonetic conditioning?
  - Can models for identifying correspondences be generalized to dozens, or even hundreds of related varieties?
  - Can borrowings be identified along with cognates?
- Virtually all NLP techniques and tools assume (and require) **consistent orthography**; surface form is the key used for looking up further information
- What if spelling differs from standard orthography? What if spelling is variable? (Note: spelling also concerns tokenization)

MDL

- Formalize as associated strings, analyze data
- Works on/produces alignments of data
- No other assumptions made

(BG)	м	и	л	
(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

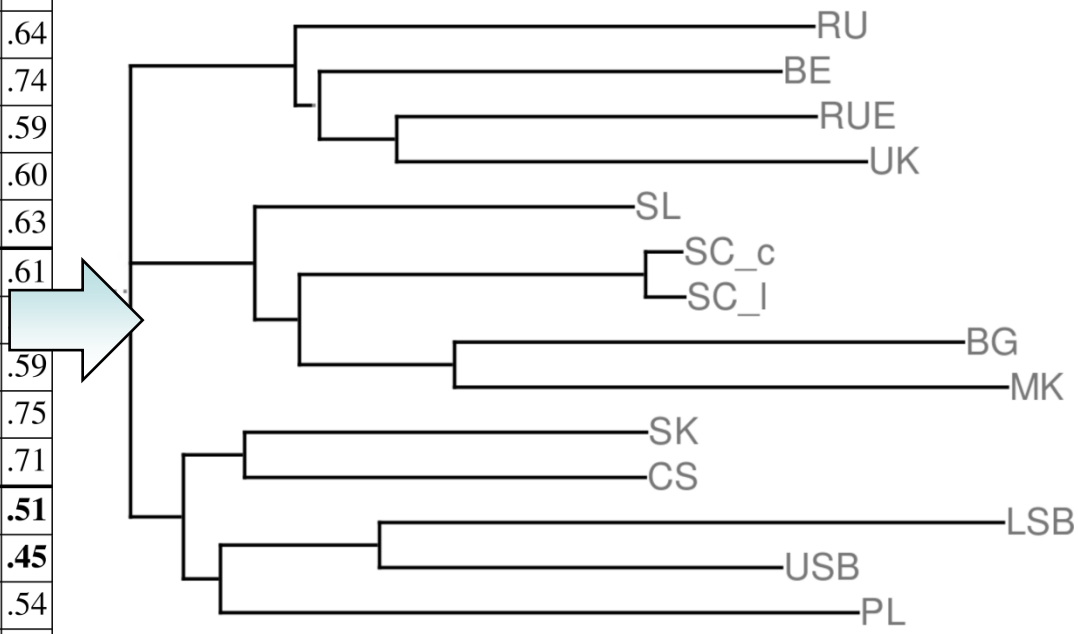
- What can we do with this?

  Objective string-level similarity:
  measures regularity and complexity of shared structure

Quantify Linguistic Similarity

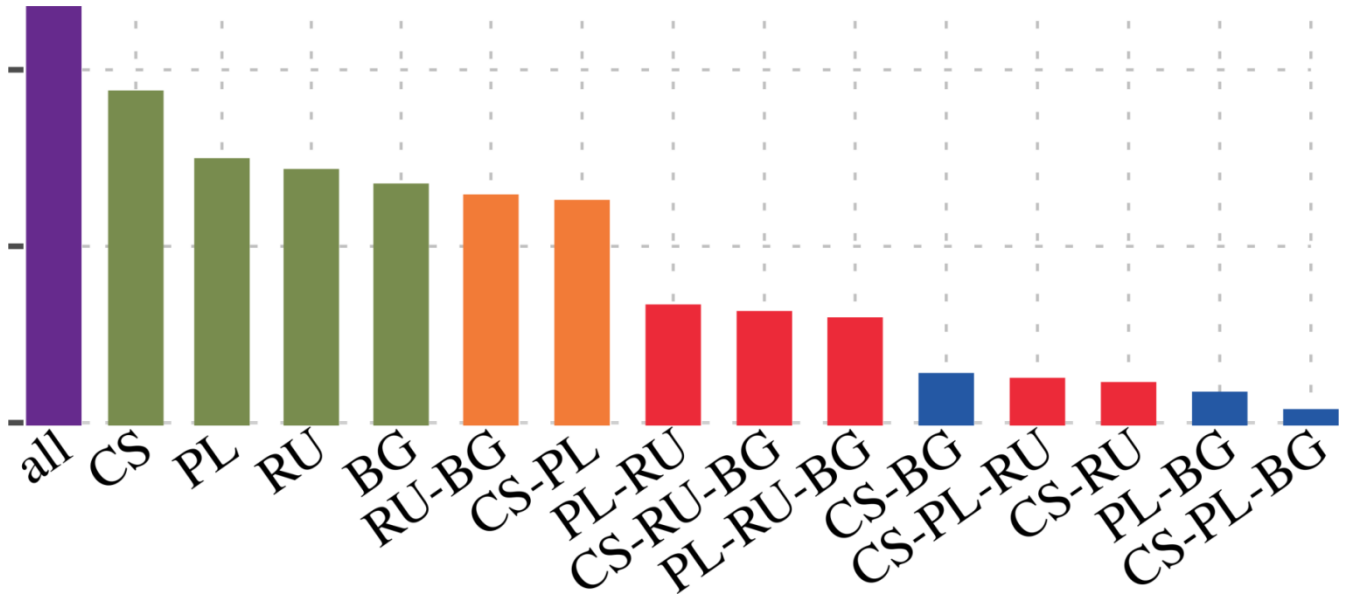
A) Phylogenetic analysis

	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



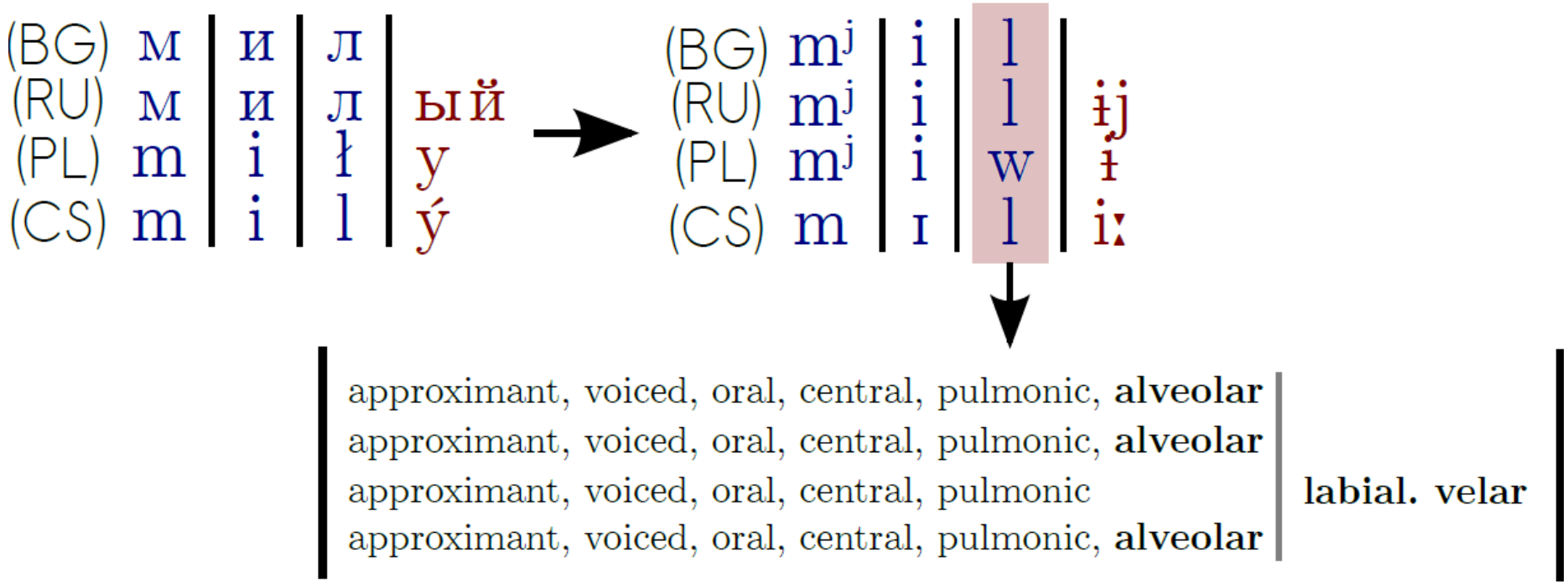
Quantify Linguistic Similarity

B) Quantify similarity within subsets of languages



Quantify Linguistic Similarity

C) Analyze both sound correspondences and sound changes



Find (And Use) Correspondences

D) Reconstruct unknown forms

(BG)	л	и		п	а
(RU)	л	и		п	а
(PL)	l		i	p	a
(CS)	l		í	p	a

(OCS)	sъč	ę	stь	je
(PL)	szcz	ę	ści	e
(CS)	št	ě	st	í
(RU)	сч	а	сть	е
(BG)	щ	а	ст	ие

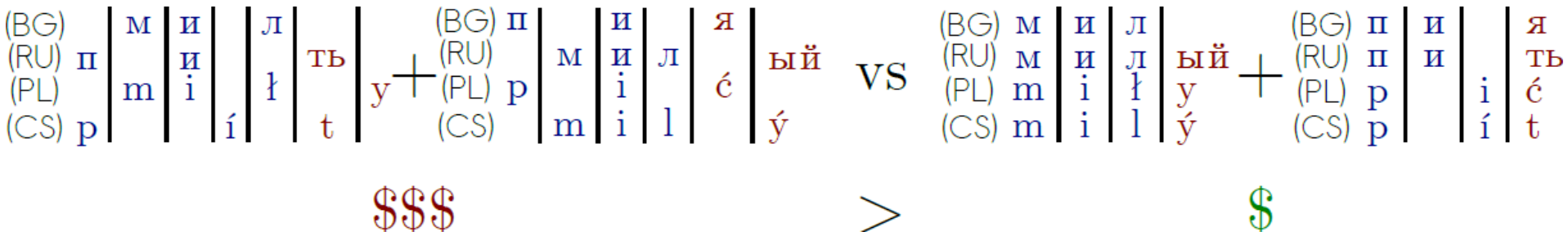
E) Analyze divergences from common spelling

(BG)	л	и		п	а
(RU)	л	и		п	а
(PL)	l		i	p	a
(CS)	l		í	p	a

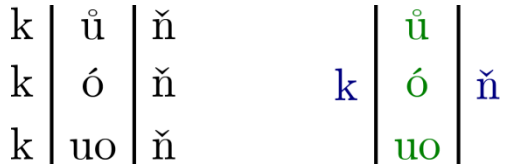
(OCS)	sъč	ę	stь	je
(PL)	szcz	ę	ści	e
(CS)	št	ě	st	í
(RU)	сч	а	сть	е
(BG)	щ	а	ст	ие

Find (And Use) Correspondences

F) Align words across languages and across time



G) Unify orthographic variants



The ultimate linguistic tool ☺ ... coming soon