A corpus-driven, usage-based approach to cross-Slavic variation using a parallel corpus

Aims, tools and first results

Ruprecht von Waldenfels

IPI PAN Warsaw & Department of Slavic languages, U Bern
waldenfels@issl.unibe.ch

IPI PAN Warsaw, 10.2.2014

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

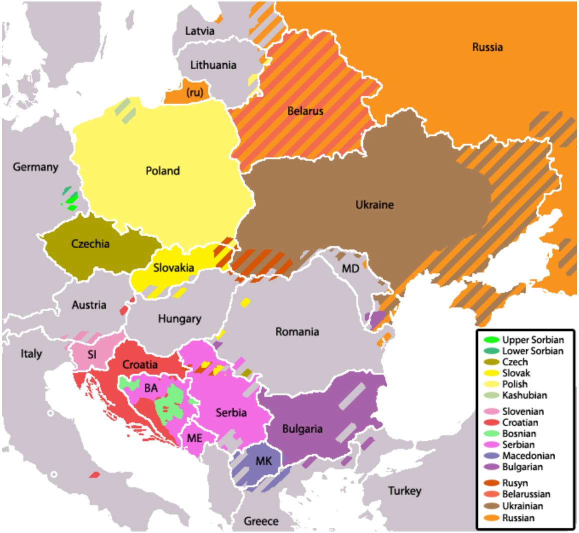
Outline

1. Introduction: Aims and methodology
    - Approaching Slavic variety from a usage point of view
    - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
    - A sample variable: Aspect in the imperative
    - Automating procedures
3. Further variables
    - Verbal prefixes
    - Nominal suffixes
    - Participles vs. finite clauses
4. Conclusions

Outline

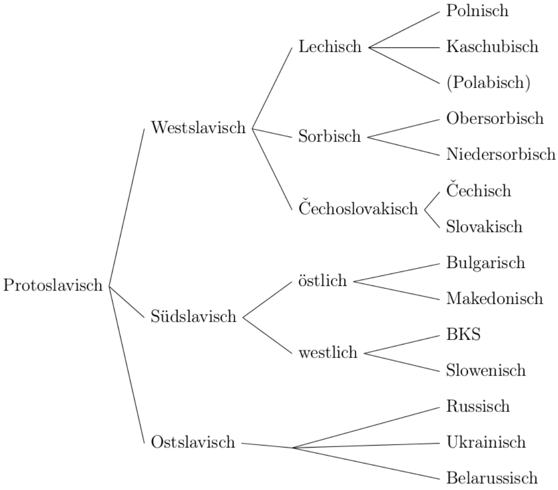
1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Slavic languages



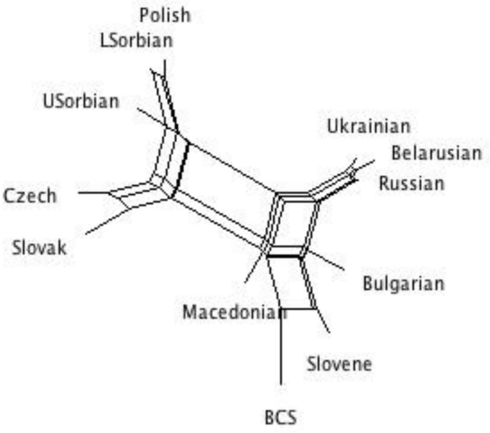
Aleksandr Vostokov (1865):
East, West and South Slavic

Slavic languages



Aleksandr Vostokov (1865):
East, West and South Slavic

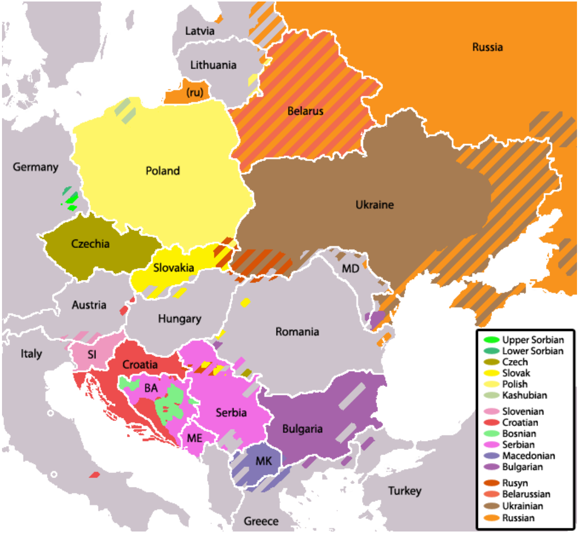
Slavic languages



Nichols & McAnallen 2012

Aleksandr Vostokov (1965):
East, West and South Slavic

Usage-based approach



Investigate difference in *use* of linguistic variables to gain insight on

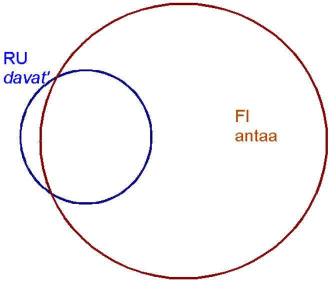
- synchronic similarity of Slavic languages (contact-induced and inherited) and their diachronic development
- the variables themselves (‘intragenealogical typology’)

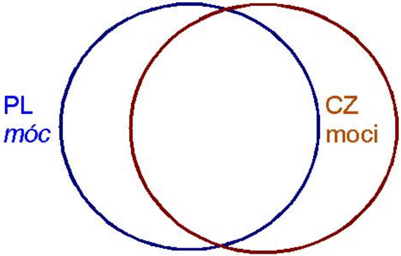
Usage-based approach

- Explore parallel corpora as a tool to compare languages bottom-up, starting from actual usage
- gain access to genealogical AND contact-based relatedness
- rigour: base our findings on in principle reproducible experiments with this data

Usage-based approach

- Explore parallel corpora as a tool to compare languages bottom-up, starting from actual usage
- gain access to genealogical AND contact-based relatedness
- rigour: base our findings on in principle reproducible experiments with this data





Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

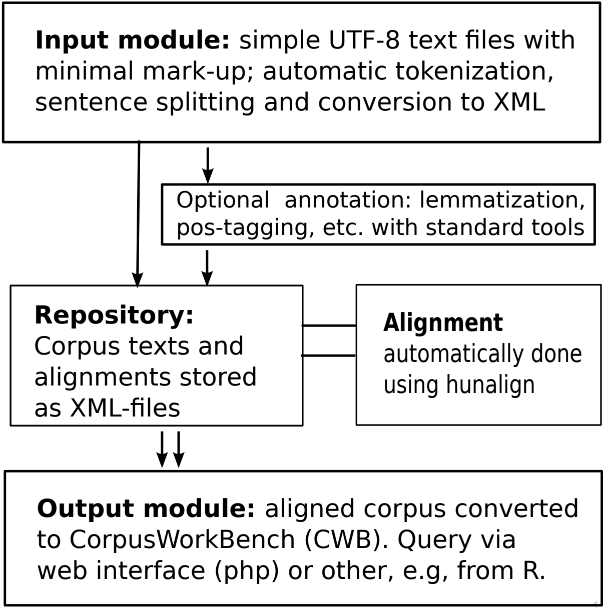
ParaSol

- U Bern and U Regensburg - developed originally jointly with Roland Meyer, then Regensburg
- Original texts and translations in Slavic languages (plus English, German, French, Greek, Baltic languages, Armenian... - ca. 30 languages) for comparative and typological linguistic research

Technical

- Framework for a quick parallel, largely language independent corpus; to be published as open source
- Corpus files in TEI-compliant XML, stand-off alignment
- Automatic alignment, linguistic annotation (POS-tagging, lemmatization) where possible (-> cooperations)
- Corpus manager OpenCorpusWorkbench, web interface in perl, php, XSLT

ParaSol: structure



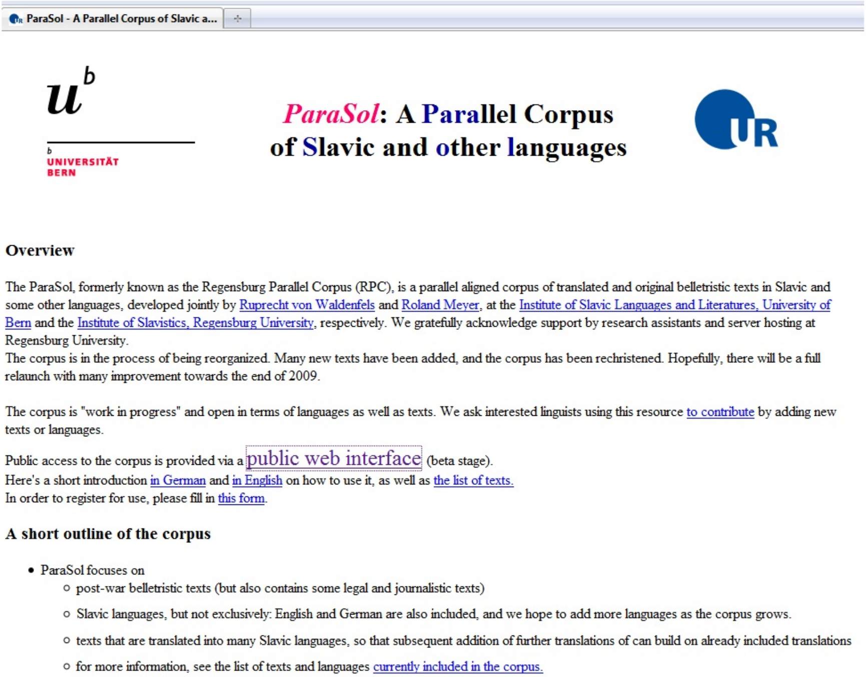
Input module: simple UTF-8 text files with minimal mark-up; automatic tokenization, sentence splitting and conversion to XML

Optional annotation: lemmatization, pos-tagging, etc. with standard tools

Repository: Corpus texts and alignments stored as XML-files

Alignment automatically done using hunalign

Output module: aligned corpus converted to CorpusWorkBench (CWB). Query via web interface (php) or other, e.g, from R.



ParaSol: A Parallel Corpus of Slavic and other languages

Overview

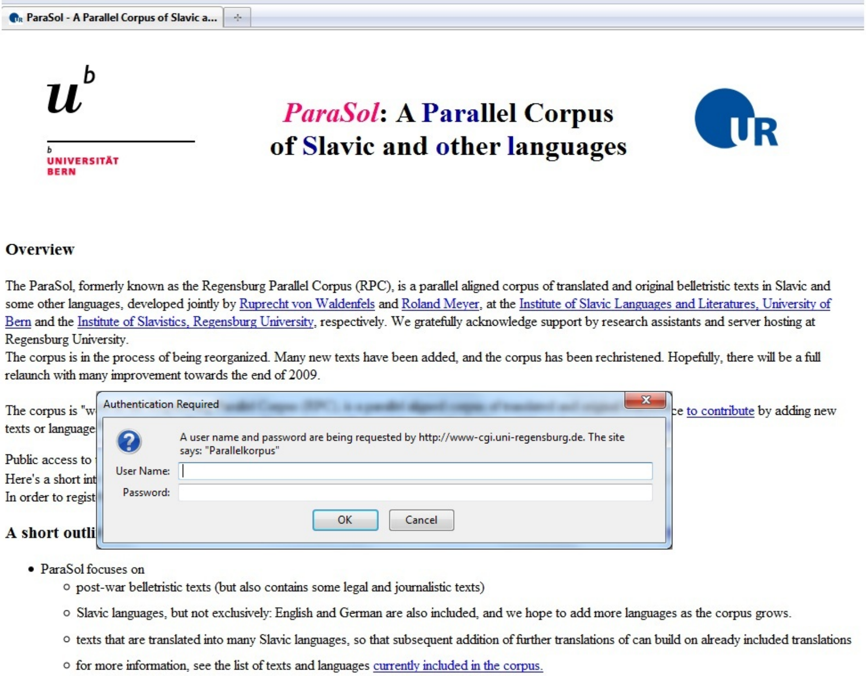
The ParaSol, formerly known as the Regensburg Parallel Corpus (RPC), is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages, developed jointly by Ruprecht von Waldenfels and Roland Meyer, at the Institute of Slavic Languages and Literatures, University of Bern and the Institute of Slavistics, Regensburg University, respectively. We gratefully acknowledge support by research assistants and server hosting at Regensburg University.
The corpus is in the process of being reorganized. Many new texts have been added, and the corpus has been rechristened. Hopefully, there will be a full relaunch with many improvement towards the end of 2009.

The corpus is "work in progress" and open in terms of languages as well as texts. We ask interested linguists using this resource to contribute by adding new texts or languages.

Public access to the corpus is provided via a public web interface (beta stage).
Here's a short introduction in German and in English on how to use it, as well as the list of texts.
In order to register for use, please fill in this form.

A short outline of the corpus

- ParaSol focuses on
  - post-war belletristic texts (but also contains some legal and journalistic texts)
  - Slavic languages, but not exclusively: English and German are also included, and we hope to add more languages as the corpus grows.
  - texts that are translated into many Slavic languages, so that subsequent addition of further translations of can build on already included translations
  - for more information, see the list of texts and languages currently included in the corpus.



ParaSol: A Parallel Corpus of Slavic and other languages

Overview

The ParaSol, formerly known as the Regensburg Parallel Corpus (RPC), is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages, developed jointly by Ruprecht von Waldenfels and Roland Meyer, at the Institute of Slavic Languages and Literatures, University of Bern and the Institute of Slavistics, Regensburg University, respectively. We gratefully acknowledge support by research assistants and server hosting at Regensburg University.
The corpus is in the process of being reorganized. Many new texts have been added, and the corpus has been rechristened. Hopefully, there will be a full relaunch with many improvement towards the end of 2009.

The corpus is "w ce to contribute by adding new texts or language

Public access to
Here's a short int
In order to regist

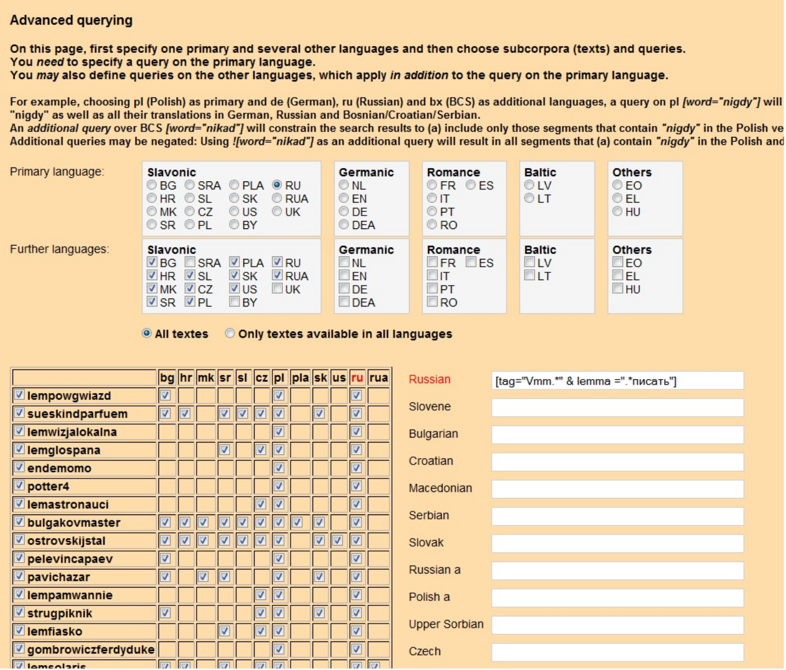
Authentication Required

A user name and password are being requested by http://www-cgi.uni-regensburg.de. The site says: "Parallelkorpus"

User Name:
Password:

A short outli

- ParaSol focuses on
  - post-war belletristic texts (but also contains some legal and journalistic texts)
  - Slavic languages, but not exclusively: English and German are also included, and we hope to add more languages as the corpus grows.
  - texts that are translated into many Slavic languages, so that subsequent addition of further translations of can build on already included translations
  - for more information, see the list of texts and languages currently included in the corpus.



Advanced querying

On this page, first specify one primary and several other languages and then choose subcorpora (texts) and queries.
You *need* to specify a query on the primary language.
You *may* also define queries on the other languages, which apply *in addition* to the query on the primary language.

For example, choosing pl (Polish) as primary and de (German), ru (Russian) and bx (BCS) as additional languages, a query on pl *[word="nigdy"]* will
"nigdy" as well as all their translations in German, Russian and Bosnian/Croatian/Serbian.
An *additional query* over BCS *[word="nikad"]* will constrain the search results to (a) include only those segments that contain *"nigdy"* in the Polish ve
Additional queries may be negated: Using *![word="nikad"]* as an additional query will result in all segments that (a) contain *"nigdy"* in the Polish and

Primary language:

Slavonic
○ BG ○ SRA ○ PLA ● RU
○ HR ○ SL ○ SK ○ RUA
○ MK ○ CZ ○ US ○ UK
○ SR ○ PL ○ BY

Germanic
○ NL
○ EN
○ DE
○ DEA

Romance
○ FR ○ ES
○ IT
○ PT
○ RO

Baltic
○ LV
○ LT

Others
○ EO
○ EL
○ HU

Further languages:

Slavonic
☑ BG □ SRA ☑ PLA ☑ RU
☑ HR ☑ SL ☑ SK ☑ RUA
☑ MK ☑ CZ ☑ US □ UK
☑ SR ☑ PL □ BY

Germanic
□ NL
□ EN
□ DE
□ DEA

Romance
□ FR □ ES
□ IT
□ PT
□ RO

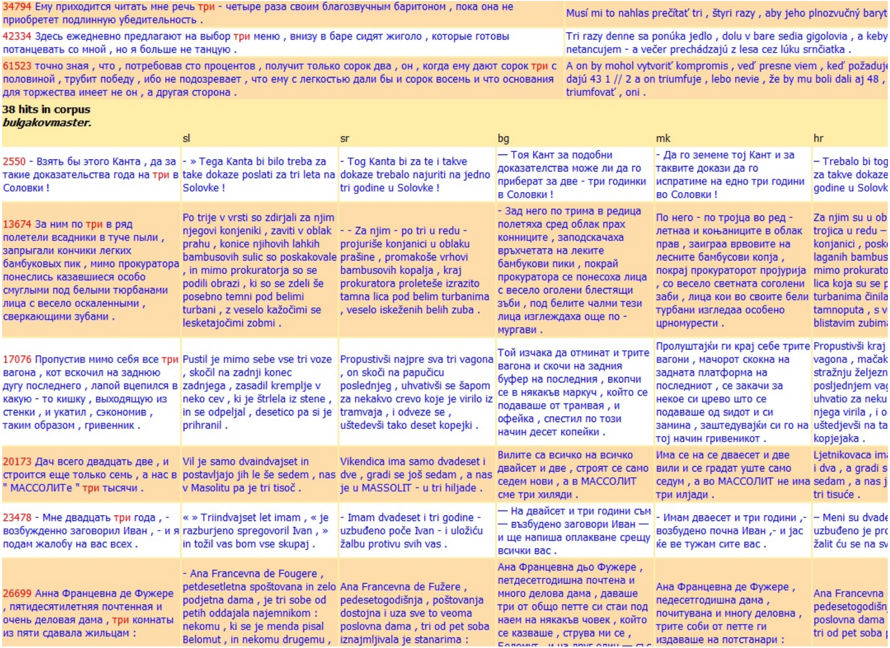
Baltic
□ LV
□ LT

Others
□ EO
□ EL
□ HU

● **All textes** ○ **Only textes available in all languages**

	bg	hr	mk	sr	sl	cz	pl	pla	sk	us	ru	rua
☑ lempowgwiazd	☑						☑				☑	
☑ sueskindparfuem	☑	☑		☑	☑	☑	☑		☑		☑	
☑ lemwizjalokalna							☑				☑	
☑ lemglospana				☑		☑	☑				☑	
☑ endemomo							☑				☑	
☑ potter4							☑				☑	
☑ lemastronauci						☑	☑				☑	
☑ bulgakovmaster	☑	☑	☑	☑	☑	☑	☑	☑	☑		☑	
☑ ostrovskijstal	☑	☑	☑	☑	☑	☑	☑		☑	☑	☑	
☑ pelevincapaev	☑						☑				☑	
☑ pavichazar	☑		☑	☑			☑		☑		☑	
☑ lempamwannie						☑	☑				☑	
☑ strugpiknik	☑					☑	☑		☑		☑	
☑ lemfiasko				☑		☑	☑				☑	
☑ gombrowiczferdyduke							☑				☑	
☑ lemsolaris	☑	☑		☑		☑	☑				☑	☑

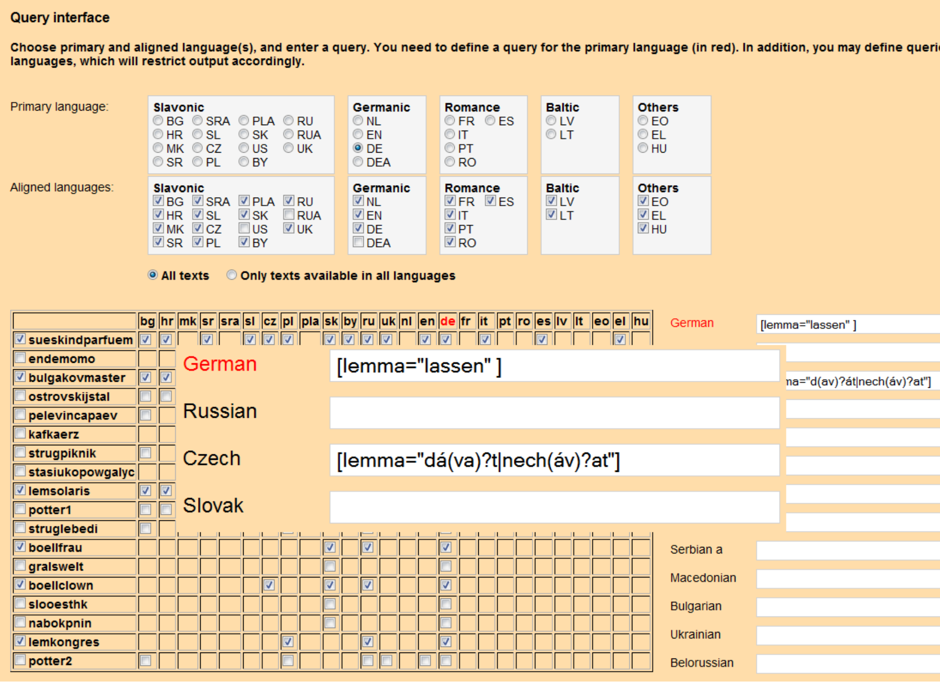
Russian	[tag="Vmm.*" & lemma =".*писать"]
Slovene	
Bulgarian	
Croatian	
Macedonian	
Serbian	
Slovak	
Russian a	
Polish a	
Upper Sorbian	
Czech	



34794 Ему приходится читать мне речь три - четыре раза своим благозвучным баритоном , пока она не приобретет подлинную убедительность .	Musí mi to nahlas prečítať tri , štyri razy , aby jeho plnozvučný baryt
42334 Здесь ежедневно предлагают на выбор три меню , внизу в баре сидят жиголо , которые готовы потанцевать со мной , но я больше не танцую .	Tri razy denne sa ponúka jedlo , dolu v bare sedia gigolovia , a keby netancujem - a večer prechádzajú z lesa cez lúku srnčiatka .
61523 точно зная , что , потребовав сто процентов , получит только сорок два , он , когда ему дают сорок три с половиной , трубит победу , ибо не подозревает , что ему с легкостью дали бы и сорок восемь и что основания для торжества имеет не он , а другая сторона .	A on by mohol vytvoriť kompromis , veď presne viem , keď požaduje dajú 43 1 // 2 a on triumfuje , lebo nevie , že by mu boli dali aj 48 , triumfovať , oni .

38 hits in corpus
bulgakovmaster.

	sl	sr	bg	mk	hr
2550 - Взять бы этого Канта , да за такие доказательства года на три в Соловки !	- » Tega Kanta bi bilo treba za take dokaze poslati za tri leta na Solovke !	- Tog Kanta bi za te i takve dokaze trebalo najuriti na jedno tri godine u Solovke !	— Тоя Кант за подобни доказателства може ли да го приберат за две - три годинки в Соловки !	- Да го земеме тој Кант и за таквите докази да го испратиме на едно три години во Соловки !	– Trebalo bi tog za takve dokaze godine u Solovk
13674 За ним по три в ряд полетели всадники в туче пыли , запрыгали кончики легких бамбуковых пик , мимо прокуратора понеслись казавшиеся особо смуглыми под белыми тюрбанами лица с весело оскаленными , сверкающими зубами .	Po trije v vrsti so zdirjali za njim njegovi konjeniki , zaviti v oblak prahu , konice njihovih lahkih bambusovih sulic so poskakovale , in mimo prokuratorja so se podili obrazi , ki so se zdeli še posebno temni pod belimi turbani , z veselo kažočimi se lesketajočimi zobmi .	- - Za njim - po tri u redu - projuriše konjanici u oblaku prašine , promakoše vrhovi bambusovih kopalja , kraj prokuratora proleteše izrazito tamna lica pod belim turbanima , veselo iskeženih belih zuba .	- Зад него по трима в редица полетяха сред облак прах конниците , заподскачаха връхчетата на леките бамбукови пики , покрай прокуратора се понесоха лица с весело оголени блестящи зъби , под белите чалми тези лица изглеждаха още по - мургави .	По него - по тројца во ред - летнаа и коњаниците в облак прав , заиграа врвовите на лесните бамбусови копја , покрај прокураторот пројурија , со весело светната соголени заби , лица кои во своите бели турбани изгледаа особено црномурести .	Za njim su u ob trojica u redu – konjanici , posk laganih bambus mimo prokurato lica koja su se p turbanima činila tamnoputa , s v blistavim zubima
17076 Пропустив мимо себя все три вагона , кот вскочил на заднюю дугу последнего , лапой вцепился в какую - то кишку , выходящую из стенки , и укатил , сэкономив , таким образом , гривенник .	Pustil je mimo sebe vse tri voze , skočil na zadnji konec zadnjega , zasadil kremplje v neko cev , ki je štrlela iz stene , in se odpeljal , desetico pa si je prihranil .	Propustivši najpre sva tri vagona , on skoči na papučicu poslednjeg , uhvativši se šapom za nekakvo crevo koje je virilo iz tramvaja , i odveze se , uštedevši tako deset kopejki .	Той изчака да отминат и трите вагона и скочи на задния буфер на последния , вкопчи се в някакъв маркуч , който се подаваше от трамвая , и офейка , спестил по този начин десет копейки .	Пролуштајќи ги крај себе трите вагони , мачорот скокна на задната платформа на последниот , се закачи за некое си цревo што се подаваше од ѕидот и си замина , заштедувајќи си го на тој начин гривеникот .	Propustivši kraj vagona , mačak stražnju željezn posljednjem vag uhvatio za neku njega virila , i o uštedjevši na ta kopjejaka .
20173 Дач всего двадцать две , и строится еще только семь , а нас в " МАССОЛИТе " три тысячи .	Vil je samo dvaindvajset in postavljajo jih le še sedem , nas v Masolitu pa je tri tisoč .	Vikendica ima samo dvadeset i dve , gradi se još sedam , a nas je u MASSOLIT - u tri hiljade .	Вилите са всичко на всичко двайсет и две , строят се само седем нови , а в МАССОЛИТ сме три хиляди .	Има се на се дваесет и две вили и се градат уште само седум , а во МАССОЛИТ не има три илјади .	Ljetnikovaca ima i dva , a gradi s sedam , a nas j tri tisuće .
23478 - Мне двадцать три года , - возбужденно заговорил Иван , - и я подам жалобу на вас всех .	« » Triindvajset let imam , « je razburjeno spregovoril Ivan , » in tožil vas bom vse skupaj .	- Imam dvadeset i tri godine - uzbuđeno poče Ivan - i uložiću žalbu protivu svih vas .	— На двайсет и три години съм — възбудено заговори Иван — и ще напиша оплакване срещу всички вас .	- Имам дваесет и три години ,- возбудено почна Иван ,- и јас ќе ве тужам сите вас .	– Meni su dvade uzbuđeno je pro žalit ću se na sv
26699 Анна Францевна де Фужере , пятидесятилетняя почтенная и очень деловая дама , три комнаты из пяти сдавала жильцам :	- Ana Francevna de Fougere , petdesetletna spoštovana in zelo podjetna dama , je tri sobe od petih oddajala najemnikom : nekomu , ki se je menda pisal Belomut , in nekomu drugemu ,	Ana Francevna de Fužere , pedesetogodišnja , poštovanja dostojna i uza sve to veoma poslovna dama , tri od pet soba iznajmljivala je stanarima :	Ана Францевна дьо Фужере , петдесетгодишна почтена и много делова дама , даваше три от общо петте си стаи под наем на някакъв човек , който се казваше , струва ми се ,	Ана Францевна де Фужере , педесетгодишна дама , почитувана и многу деловна , трите соби от петте ги издаваше на потстанари :	Ana Francevna pedesetogodišn poslovna dama tri od pet soba



Query interface

Choose primary and aligned language(s), and enter a query. You need to define a query for the primary language (in red). In addition, you may define queri languages, which will restrict output accordingly.

Primary language:

Slavonic
○ BG ○ SRA ○ PLA ○ RU
○ HR ○ SL ○ SK ○ RUA
○ MK ○ CZ ○ US ○ UK
○ SR ○ PL ○ BY

Germanic
○ NL
○ EN
● DE
○ DEA

Romance
○ FR ○ ES
○ IT
○ PT
○ RO

Baltic
○ LV
○ LT

Others
○ EO
○ EL
○ HU

Aligned languages:

Slavonic
☑ BG ☑ SRA ☑ PLA ☑ RU
☑ HR ☑ SL ☑ SK □ RUA
☑ MK ☑ CZ □ US ☑ UK
☑ SR ☑ PL ☑ BY

Germanic
☑ NL
☑ EN
☑ DE
□ DEA

Romance
☑ FR ☑ ES
☑ IT
☑ PT
☑ RO

Baltic
☑ LV
☑ LT

Others
☑ EO
☑ EL
☑ HU

● **All texts** ○ **Only texts available in all languages**

	bg	hr	mk	sr	sra	sl	cz	pl	pla	sk	by	ru	uk	nl	en	de	fr	it	pt	ro	es	lv	lt	eo	el	hu
☑ sueskindparfuem	☑	☑		☑		☑	☑	☑		☑	☑	☑	☑		☑	☑		☑			☑				☑	
□ endemomo																										
☑ bulgakovmaster	☑	☑																								
□ ostrovskijstal	□	□																								
□ pelevincapaev	□																									
□ kafkaerz																										
□ strugpiknik	□																									
□ stasiukopowgalyc																										
☑ lemsolaris	☑	☑																								
□ potter1	□	□																								
□ struglebedi	□																									
☑ boellfrau										☑		☑				☑										
□ gralswelt										□						□										
☑ boellclown							☑			☑		☑				☑										
□ slooesthk										□						□										
□ nabokpnin										□						□										
☑ lemkongres								☑				☑				☑										
□ potter2	□							□				□	□		□	□										

German [lemma="lassen"]

Russian

Czech [lemma="dá(va)?t|nech(áv)?at"]

Slovak

German [lemma="lassen"]

na="d(av)?át|nech(áv)?at"]

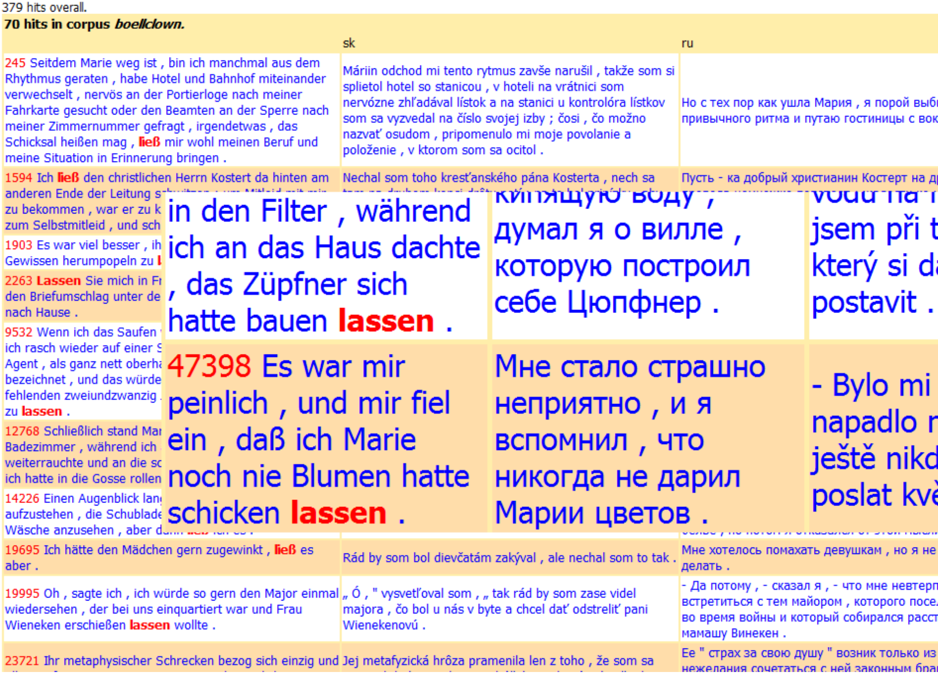
Serbian a

Macedonian

Bulgarian

Ukrainian

Belorussian



379 hits overall.

70 hits in corpus *boellclown*.

	sk	ru
245 Seitdem Marie weg ist , bin ich manchmal aus dem Rhythmus geraten , habe Hotel und Bahnhof miteinander verwechselt , nervös an der Portierloge nach meiner Fahrkarte gesucht oder den Beamten an der Sperre nach meiner Zimmernummer gefragt , irgendetwas , das Schicksal heißen mag , **ließ** mir wohl meinen Beruf und meine Situation in Erinnerung bringen .	Máriin odchod mi tento rytmus zavše narušil , takže som si splietol hotel so stanicou , v hoteli na vrátnici som nervózne zhľadával lístok a na stanici u kontrolóra lístkov som sa vyzvedal na číslo svojej izby ; čosi , čo možno nazvať osudom , pripomenulo mi moje povolanie a položenie , v ktorom som sa ocitol .	Но с тех пор как ушла Мария , я порой выб привычного ритма и путаю гостиницы с вок
1594 Ich **ließ** den christlichen Herrn Kostert da hinten am anderen Ende der Leitung s zu bekommen , war er zu k zum Selbstmitleid , und sch	Nechal som toho kresťanského pána Kosterta , nech sa	Пусть - ка добрый христианин Костерт на д
1903 Es war viel besser , ih Gewissen herumpopeln zu **l**		
2263 **Lassen** Sie mich in Fr den Briefumschlag unter de nach Hause .		
9532 Wenn ich das Saufen ich rasch wieder auf einer S Agent , als ganz nett oberha bezeichnet , und das würde fehlenden zweiundzwanzig . zu **lassen** .		
12768 Schließlich stand Mar Badezimmer , während ich weiterrauchte und an die sc ich hatte in die Gosse rollen		
14226 Einen Augenblick lang aufzustehen , die Schublade Wäsche anzusehen , aber d		
19695 Ich hätte den Mädchen gern zugewinkt , **ließ** es aber .	Rád by som bol dievčatám zakývať , ale nechal som to tak .	Мне хотелось помахать девушкам , но я не делать .
19995 Oh , sagte ich , ich würde so gern den Major einmal wiedersehen , der bei uns einquartiert war und Frau Wieneken erschießen **lassen** wollte .	„ Ó , " vysvetľoval som , „ tak rád by som zase videl majora , čo bol u nás v byte a chcel dať odstreliť pani Wienekenovú .	- Да потому , - сказал я , - что мне невтерп встретиться с тем майором , которого посел во время войны и который собирался расст мамашу Винекен .
23721 Ihr metaphysischer Schrecken bezog sich einzig und	Jej metafyzická hrôza pramenila len z toho , že som sa	Ее " страх за свою душу " возник только из нежелания сочетаться с ней законным бра

in den Filter , während ich an das Haus dachte , das Züpfner sich hatte bauen **lassen** .	кипящую воду , думал я о вилле , которую построил себе Цюпфнер .	vodu na jsem při t který si d postavit .
47398 Es war mir peinlich , und mir fiel ein , daß ich Marie noch nie Blumen hatte schicken **lassen** .	Мне стало страшно неприятно , и я вспомнил , что никогда не дарил Марии цветов .	- Bylo mi napadlo ještě nikd poslat kvě

State of corpus

LNG	in full	tokens	lemmas	tags
BG	Bulgarian	2 002 697	43 280	y
BY	Belarusian	482 467	24 131	y
CZ	Czech	1 629 868	47 166	y
DA	Danish	100 448	0	n
DE	German	2 006 781	64 112	y
EE	Estonian	287 948	23 031	y
EL	Greek	600 594	40 141	y
EN	English	814 289	19 886	y
EO	Esperanto	152 660	0	n
ES	Spanish	476 301	21 414	y
FI	Finnish	174 204	0	n
FR	French	448 612	10 870	y
HR	Croatian	899 466	43 263	y
HU	Hungarian	146 505	0	n
HY	Armenian	240 815	7 873	y
IT	Italian	478 315	19 472	y
LT	Lithuanian	280 675	8 001	y
LV	Latvian	147 906	0	n
MK	Macedonian	1 045 873	43 646	y
NL	Dutch	728 061	4 836	y
NO	Norwegian	334 948	13 788	y
PL	Polish	3 396 673	66 492	y
PT	Portuguese	380 659	10 961	y
RO	Romanian	398 088	15 164	y
RU	Russian	3 637 357	78 997	y
SK	Slovak	1 457 925	51 010	y
SL	Slovene	1 132 839	36 229	y
SR	Serbian	1 324 929	42 602	y
SV	Swedish	314 759	0	n
UK	Ukrainian	1 017 054	33 562	y
US	Upper Sorbian	73 266	0	n

32 languages, >400 pairs, 25 mio token
In all major Slavic languages:

- Michail Bulgakov: Master i Margarita
- Stanisław Lem: Solaris
- Umberto Eco: Il nome della rosa
- Patrick Süskind: Das Parfüm
- Nikolaj Ostrovskij: Kak zakaljalos' stal'
- Joanne K. Rowling: Harry Potter and the Sorcerer's Stone
- Ivo Andrić: Na Drini ćuprija (still incomplete)
- Milan Kundera: Nesnesitelná lehkost bytí (still incomplete)

`http://parasol.unibe.ch`

texts	BG	BY	CZ	DA	DE	DEa	EE	EL	EN	EO	ES	FI	FR	HR	HU	HY	IT	LT	LV	MK	NL	NO	PL	PLa	PT	RO	RU	RUa	SK	SL
AndricDrina		BY	CZ																				PL				RU			SL
BoellClown	BG		CZ		DE																						RU		SK	
BoellFrau					DE																						RU		SK	
BulgakovMaster	BG	BY	CZ		DE		EE	EL	EN	EO	ES		FR	HR	HU	HY	IT	LT	LV	MK	NL		PL	PLa	PT	RO	RU		SK	SL
DimkovskaSkrienaKamera																				MK									SK	SL
DimovDusi	BG																												SK	
EUVerf					DE																								SK	
EcoRosa	BG		CZ		DE			EL	EN		ES	FI		HR			IT			MK	NL	NO	PL				RU		SK	SL
EndeMomo					DE																		PL				RU			
GombrowiczFerdydurke			CZ																		NL		PL				RU			SL
GralsWelt					DE																								SK	
KafkaErz					DE																								SK	
KunderaLehkost	BG		CZ	DA	DE				EN									LT		MK		NO	PL				RU			
LemAstronauci			CZ																				PL				RU			
LemFiasko	BG		CZ											HR									PL				RU			
LemGlosPana			CZ																				PL				RU			
LemKatar	BG																						PL				RU			
LemKongres					DE																		PL				RU			
LemPamWannie			CZ																				PL				RU			
LemPokoj			CZ																				PL				RU			
LemPowGwiazd	BG		CZ																				PL				RU			
LemSolaris	BG	BY	CZ		DE		EE							HR						MK			PL				RU	RUa	SK	SL
LemWizjaLokalna	BG																						PL				RU			
NabokovPnin					DE	DEa																					RU	RUa	SK	
OstrovskijStal	BG	BY	CZ		DE									HR						MK			PL				RU		SK	SL
PavicHazar	BG																			MK			PL				RU		SK	
PelevinCapaev	BG				DE																		PL				RU			
Potter1	BG		CZ		DE			EL	EN				FR	HR		HY				MK	NL		PL		PT	RO	RU	RUa	SK	SL
Potter2	BG				DE				EN				FR								NL		PL		PT	RO	RU	RUa		
Potter3	BG																						PL				RU	RUa		
Potter4																							PL				RU			
Potter5																							PL				RU			
SloOestHK					DE																								SK	
StarovaVremeto																				MK										
StasiukDziewiec														HR									PL				RU			SL
StasiukOpowGalyc					DE																		PL							
StrugLebedi	BG				DE																		PL				RU			
StrugPiknik	BG		CZ		DE																		PL				RU		SK	
SueskindParfuem	BG	BY	CZ		DE		EE	EL	EN		ES			HR			IT	LT		MK			PL				RU		SK	SL

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages

2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures

3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses

4. Conclusions

An example feature

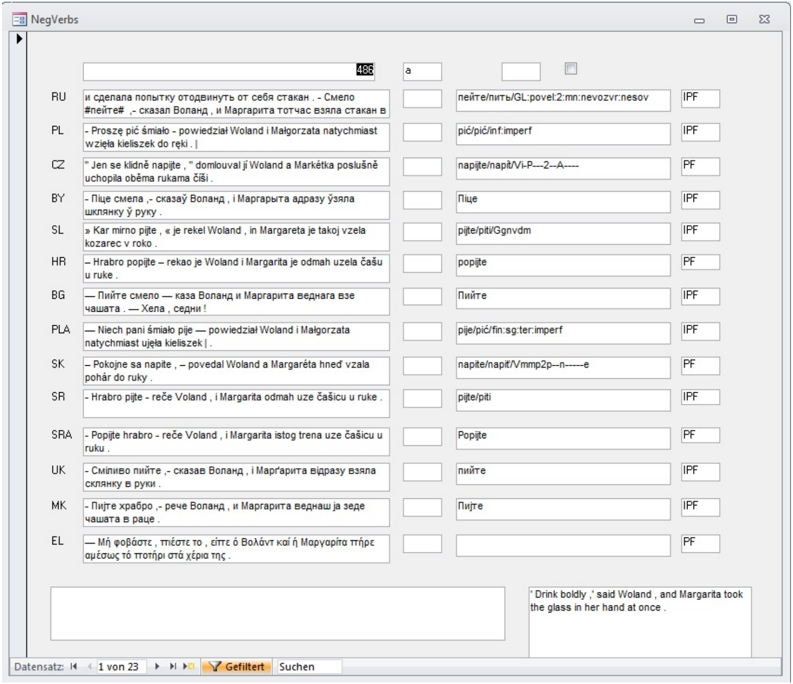
For example, aspectual opposition in the imperative:

> In some Slavic languages, an imperfective verb form may be used instead of an (actionally) expected perfective imperative form to introduce a pragmatic effect; e.g., express a *contextually expected request*.

(1) Vključajte (IPF) televizor, uže sem' časov. Peredača načinaetsja.
Turn on the TV, it's already 7 o'clock - the show starts.

(2) Vključite (PF) televizor, segodnja interesnaja peredača.
Turn on the TV, there's something interesting showing.

(Rassudova 1982; Benacchio 2010 on variation in Slavic)



	486	a		☐
RU	и сделала попытку отодвинуть от себя стакан . - Смело #пейте# ,- сказал Воланд , и Маргарита тотчас взяла стакан в		пейте/пить/GL:povel:2:mn:nevozvr:nesov	IPF
PL	- Proszę pić śmiało - powiedział Woland i Małgorzata natychmiast wzięła kieliszek do ręki . |		pić/pić/inf:imperf	IPF
CZ	" Jen se klidně napijte , " domlouval jí Woland a Markétka poslušně uchopila oběma rukama číši .		napijte/napít/Vi-P---2--A----	PF
BY	- Піце смела ,- сказаў Воланд , і Маргарыта адразу ўзяла шклянку ў руку .		Піце	IPF
SL	» Kar mirno pijte , « je rekel Woland , in Margareta je takoj vzela kozarec v roko .		pijte/piti/Ggnvdm	IPF
HR	– Hrabro popijte – rekao je Woland i Margarita je odmah uzela čašu u ruke .		popijte	PF
BG	— Пийте смело — каза Воланд и Маргарита веднага взе чашата . — Хела , седни !		Пийте	IPF
PLA	— Niech pani śmiało pije — powiedział Woland i Małgorzata natychmiast ujęła kieliszek | .		pije/pić/fin:sg:ter:imperf	IPF
SK	– Pokojne sa napite , – povedal Woland a Margaréta hneď vzala pohár do ruky .		napite/napiť/Vmmp2p--n-----e	PF
SR	- Hrabro pijte - reče Voland , i Margarita odmah uze čašicu u ruke .		pijte/piti	IPF
SRA	- Popijte hrabro - reče Voland , i Margarita istog trena uze čašicu u ruku .		Popijte	PF
UK	- Сміливо пийте ,- сказав Воланд , і Марґарита відразу взяла склянку в руки .		пийте	IPF
MK	- Пијте храбро ,- рече Воланд , и Маргарита веднаш ја зеде чашата в раце .		Пијте	IPF
EL	— Μή φοβάστε , πιέστε το , είπε ό Βολάντ καί ή Μαργαρίτα πήρε αμέσως τό ποτήρι στά χέρια της .			PF

' Drink boldly ,' said Woland , and Margarita took the glass in her hand at once .

No variation

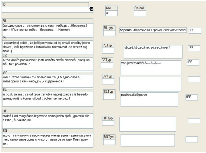
(3) Perfective across Slavic:
'*Take him outside* for a moment, explain to him how I ought to be spoken to.'

RU Vyvedite^{pfv} ego otsjuda na minutu, ob''jasnite emu, kak nado razgovarivat' so mnoj.

PL Wyprowadź^{pfv} go stąd na chwilę i wyjaśnij mu, jak należy się do mnie zwracać.

CZ Odveď^{pfv} ho a vysvětli mu, jak se mnou má mluvit.

SL Odpeljite^{pfv} ga za trenutek od tod in mu pojasnite, kako je treba govoriti z menoj.



ID

Alle
a

Default

RU
бы одно слово , заговоришь с кем - нибудь , #берегись# меня ! Повторяю тебе : - берегись . - И [illegible]

RUtyp
берегись/беречься/GL:povel:2:ed:vozvr:neso
IPF

PL
I zapamiętaj sobie , że jeśli powiesz od tej chwili choćby jedno słowo , jeśli będziesz z kimkolwiek rozmawiał - to strzeż się mnie ! |

PLTyp
strzeż/strzec/impt:sg:sec:imperf
IPF

CZ
A teď dobře poslouchej : jestli od této chvíle hlesneš , varuj se mě , to ti povídám ! "

CZTyp
varuj/varovat/Vi-S---2--A----
IPF

BY
калі з гэтае хвіліны ты прамовіш хаця б адно слова , загаворыш з кім - небудзь ,- сцеражыся !

BYTyp

SL
In poslušaj me : če od tega trenutka naprej izrečeš le besedo , spregovoriš s komer si bodi , potem se me pazi !

SLTyp
pazi/paziti/Ggnvde
IPF

HR
budeš li od ovog časa izgovorio samo jednu riječ , govorio bilo s kime , čuvaj me se !

HRTyp

BG
ако от тази минута произнесеш макар една - единичка дума , ако само заговориш с някого , пази се от мен Повтарям ти :

BGTyp

No variation

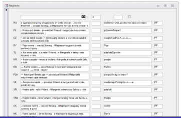
(4) Imperfective across Slavic:
’if from this moment on you say even one word, if you speak to anyone at all, *beware of me*!’

RU I slušaj menja: esli s ėtoj minuty ty proizneseš’ chotja by odno slovo, zagovoriš’ s kem - nibud’, beregis’^{ipfv} menja!

PL I zapamiętaj sobie, że jeśli powiesz od tej chwili choćby jedno słowo, jeśli będziesz z kimkolwiek rozmawiał - to strzeż się^{ipfv} mnie! |

CZ A teď dobře poslouchej: jestli od této chvíle hlesneš, varuj^{ipfv} se mě, to ti povídám! "

SL In poslušaj me: če od tega trenutka naprej izrečeš le besedo, spregovoriš s komer si bodi, potem se me pazi^{ipfv}!



Variation

(5) ’ [. . .] did you ever say anything about the great Caesar? *Answer!* Did you?’

RU	Otvečaj^{ipfv}!
PL	Odpowiadaj^{ipfv}!
CZ	Odpověz^{pfv}!
SL	Odgovori^{pfv}!
UK	Vidpovidaj^{ipfv}!
BY	Adkazvaj^{ipfv}!
BG	Otgovarjaj^{ipfv}!
MK	Odgovaraj^{ipfv}!
SR	Odgovaraj^{ipfv}!
SR2	Odgovaraj^{ipfv}!
HR	Odgovori^{pfv}!
SK	Odpovedz^{pfv}!
PL2	Odpowiadaj^{ipfv}!

Variation

(6) '*Repeat it* a third time, High Priest.'

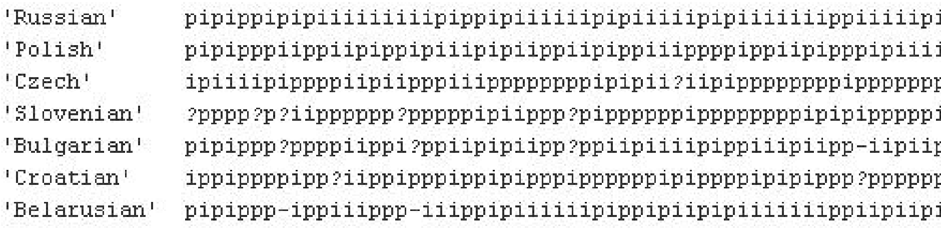
RU Pervosvjaščennik, povtori^{pfv} v tretij raz.
PL Arcykapłanie, potwórz^{pfv} to po raz trzeci.
CZ Opakuj^{ipfv} to potřetí, velekněže.
SL Veliki duhoven, ponovi^{pfv} mi to še tretjič.

Variation

(6) '*Repeat it* a third time, High Priest.'

RU Pervosvjaščennik, povtori^{pfv} v tretij raz.
PL Arcykapłanie, potwórz^{pfv} to po raz trzeci.
CZ Opakuj^{ipfv} to potřetí, velekněže.
SL Veliki duhoven, ponovi^{pfv} mi to še tretjič.
BY Peršasvjatar, paŭtary^{pfv} trėci raz!
UK Pervosvjaščeniku, povtori^{pfv} vtretje.
UK Pervosvjaščenyku, povtory^{pfv} vtretje.
BG Părvosvešteniko, potvărdi^{pfv} go i treti păt.
MK Prvosvešteniku, povtori^{pfv} go toa i po tret pat.
SR Prvosvešteniče, ponovi^{pfv} i treći put.
SR2 Prvo sveštениče, ponovi^{pfv} i treći put.
HR Prvosvećeniče, ponovi^{pfv} treći put.
SK Veľkňaz, zopakuj^{pfv} to tretí raz.
PL2 Powtórz^{pfv} to po raz trzeci, arcykapłanie.

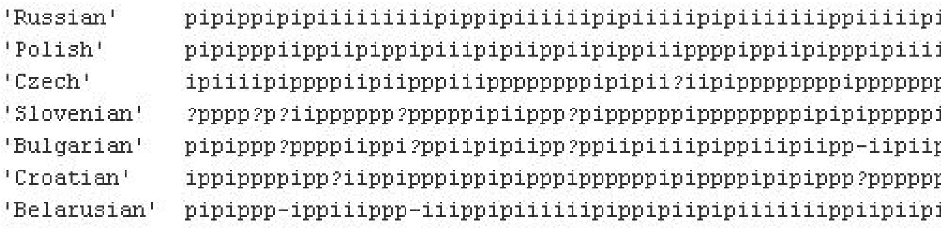
Matrix with binary values



```
'Russian'     pipippipipiiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiippppippiipipppipiiii
'Czech'       ipiiiipipppiipiippiiiippppppppipipii?iipipppppppppippppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipiippp?pipppppppippppppppipipippppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippippppipp?iippipppippipipppipppppipipppipipippp?pppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiiippiipiipi
```

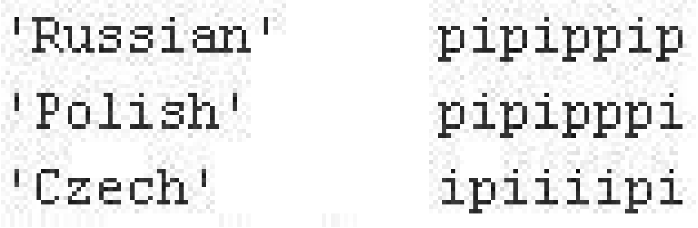
Hamming distances: for each pair, take proportion of differing symbols

Matrix with binary values



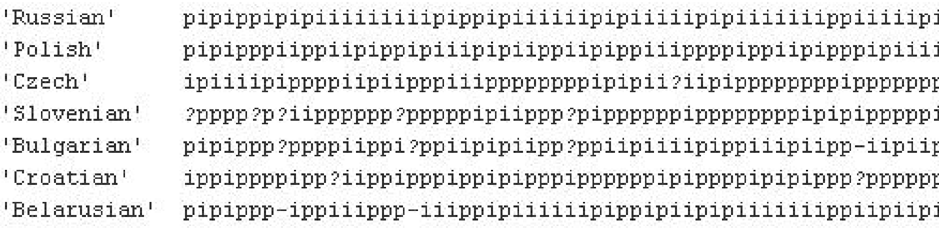
```
'Russian'    pipippipipiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'     pipipppiippiipippipiiipipiippiipippiiippppippiipipppipiiii
'Czech'      ipiiiipippppiipiipppiiippppppppipipii?iipippppppppippppppp
'Slovenian'  ?pppp?p?iipppppp?pppppipiippp?pippppppipppppppppipipipppppi
'Bulgarian'  pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'   ippippppipp?iippipppippipipppippppppipippppipipippp?ppppppp
'Belarusian' pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiiippiipiipi
```

Hamming distances: for each pair, take proportion of differing symbols



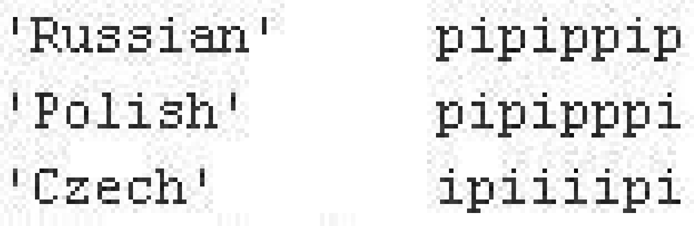
```
'Russian'    pipippip
'Polish'     pipipppi
'Czech'      ipiiiipi
```

Matrix with binary values



```
'Russian'    pipippipipiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'     pipipppiippiipippipiiipipiippiipippiiipppippiipipppipiiii
'Czech'      ipiiiipippppiipiipppiiippppppppipipii?iipippppppppipppppp
'Slovenian'  ?pppp?p?iipppppp?pppppipiippp?pippppppipppppppipipipppppi
'Bulgarian'  pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'   ippipppppipp?iippipppippipipppipppppipipppipipipppp?pppppp
'Belarusian' pipippp-ippiiippp-iiippipiiiiipippipiipipiiiiiiippiipiipi
```

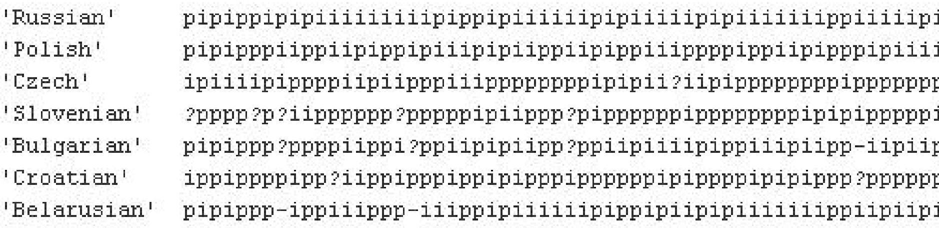
Hamming distances: for each pair, take proportion of differing symbols



```
'Russian'    pipippip
'Polish'     pipipppi
'Czech'      ipiiiipi
```

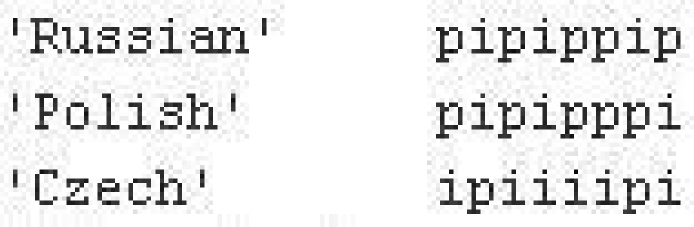
RU-PL: 2/9 =0.222

Matrix with binary values



```
'Russian'    pipippipipiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'     pipipppiippiipippipiiipipiippiipippiiippppippiipipppipiiii
'Czech'      ipiiiipipppppiipiippppiiippppppppipipii?iipipppppppppipppppppp
'Slovenian'  ?pppp?p?iippppppp?pppppipiippp?pipppppppipppppppppipipipppppi
'Bulgarian'  pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'   ippippppipp?iippipppippipipppippppppipipppppipipippp?pppppp
'Belarusian' pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiiippiipiipi
```

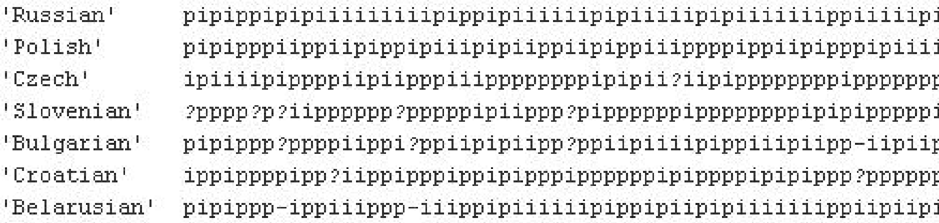
Hamming distances: for each pair, take proportion of differing symbols



```
'Russian'    pipippip
'Polish'     pipipppi
'Czech'      ipiiiipi
```

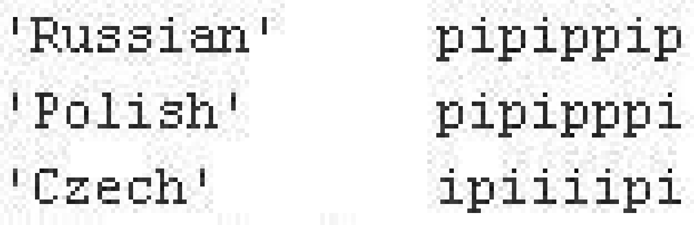
RU-PL: 2/9 =0.222   PL-CZ: 5/9 =0.555

Matrix with binary values



```
'Russian'    pipippipipiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'     pipipppiippiipippipiiipipiippiipippiiippppippiipipppipiiii
'Czech'      ipiiiipipppiipiippiiipppppppipipii?iipipppppppipppppppp
'Slovenian'  ?pppp?p?iipppppp?pppppipiippp?pippppppippppppppipipippppppi
'Bulgarian'  pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'   ippippppipp?iippipppippipippipppppipipppipipippp?pppppp
'Belarusian' pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiippiipiipi
```

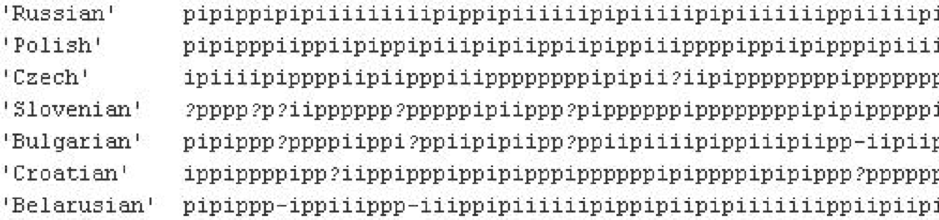
Hamming distances: for each pair, take proportion of differing symbols



```
'Russian'    pipippip
'Polish'     pipipppi
'Czech'      ipiiiipi
```

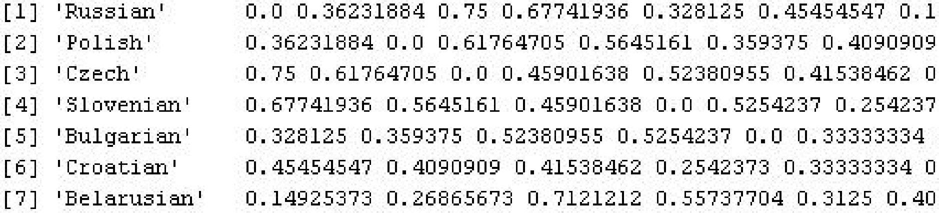
RU-PL: 2/9 =0.222   PL-CZ: 5/9 =0.555   CZ-RU: 7/9 =0.777

Matrix with binary values



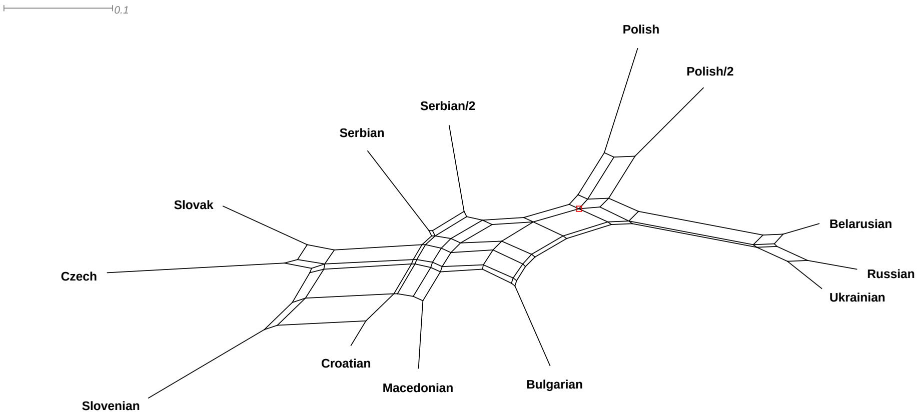
```
'Russian'     pipippipipiiiiiiiiipippipiiiiiipipiiiiipipiiiiiiippiiiiipi
'Polish'      pipipppiippiipippipiiipipiippiipippiiipppippiipipppipiiii
'Czech'       ipiiiipippppiipiipppiiippppppppipipii?iipippppppppippppppp
'Slovenian'   ?pppp?p?iipppppp?pppppipiippp?pipppppppipppppppppipipipppppi
'Bulgarian'   pipippp?ppppiippi?ppiipipiipp?ppiipiiiipippiiipiipp-iipiip
'Croatian'    ippippppipp?iippipppippipipppipppppipipppipipippp?ppppppp
'Belarusian'  pipippp-ippiiippp-iiippipiiiiiipippipiipipiiiiiiippiipiipi
```

Table of distances:



```
[1] 'Russian'       0.0 0.36231884 0.75 0.67741936 0.328125 0.45454547 0.1
[2] 'Polish'        0.36231884 0.0 0.61764705 0.5645161 0.359375 0.4090909
[3] 'Czech'         0.75 0.61764705 0.0 0.45901638 0.52380955 0.41538462 0
[4] 'Slovenian'     0.67741936 0.5645161 0.45901638 0.0 0.5254237 0.254237
[5] 'Bulgarian'     0.328125 0.359375 0.52380955 0.5254237 0.0 0.33333334
[6] 'Croatian'      0.45454547 0.4090909 0.41538462 0.2542373 0.33333334 0
[7] 'Belarusian'    0.14925373 0.26865673 0.7121212 0.55737704 0.3125 0.40
```

Visualizing distances with NeighborNet



Neighbor net of Bulgakov's *Master i Margarita*

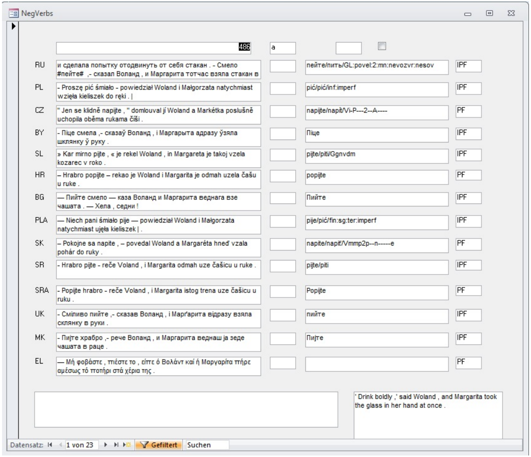
Aspect variable: Results

- Eastern and Western Group as found in the literature (Barentsen 1995, Dickey 2000, Benacchio 2010)
- Interesting to compare clustering across many domains - independent evidence for or against aspectual prototype relevant for all domains in a language
- comparison across more contexts desirable

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Next step: more comprehensive, automated measurements



	486	a		☐
RU	и сделала попытку отодвинуть от себя стакан . - Смело #пейте# ,- сказал Воланд , и Маргарита тотчас взяла стакан в		пейте/пить/GL:povel:2:mn:nevozvr:nesov	IPF
PL	- Proszę pić śmiało - powiedział Woland i Małgorzata natychmiast wzięła kieliszek do ręki . |		pić/pić/inf:imperf	IPF
CZ	" Jen se klidně napijte , " domlouval jí Woland a Markétka poslušně uchopila oběma rukama číši .		napijte/napít/Vi-P---2--A----	PF
BY	- Піце смела ,- сказаў Воланд , і Маргарыта адразу ўзяла шклянку ў руку .		Піце	IPF
SL	» Kar mirno pijte , « je rekel Woland , in Margareta je takoj vzela kozarec v roko .		pijte/piti/Ggnvdm	IPF
HR	– Hrabro popijte – rekao je Woland i Margarita je odmah uzela čašu u ruke .		popijte	PF
BG	— Пийте смело — каза Воланд и Маргарита веднага взе чашата . — Хела , седни !		Пийте	IPF
PLA	— Niech pani śmiało pije — powiedział Woland i Małgorzata natychmiast ujęła kieliszek | .		pije/pić/fin:sg:ter:imperf	IPF
SK	– Pokojne sa napite , – povedal Woland a Margaréta hneď vzala pohár do ruky .		napite/napiť/Vmmp2p--n-----e	PF
SR	- Hrabro pijte - reče Voland , i Margarita odmah uze čašicu u ruke .		pijte/piti	IPF
SRA	- Popijte hrabro - reče Voland , i Margarita istog trena uze čašicu u ruku .		Popijte	PF
UK	- Сміливо пийте ,- сказав Воланд , і Марґарита відразу взяла склянку в руки .		пийте	IPF
MK	- Пијте храбро ,- рече Воланд , и Маргарита веднаш ја зеде чашата в раце .		Пијте	IPF
EL	— Μή φοβάστε , πιέστε το , είπε ὁ Βολάντ καί ἡ Μαργαρίτα πήρε αμέσως τό ποτήρι στά χέρια της .			PF

' Drink boldly ,' said Woland , and Margarita took the glass in her hand at once .

necessary:

- tagging for *all* languages (Macedonian, Sorbian!)
- word alignment (→ Uplug, Tiedemann 2004)
- operationalization of variables

→ allows to include more variables, more texts; but introduces more noise

Setup

Data format

- word aligned texts held in large XML files / CWB Corpus
- variables operationalized with regular expressions on word forms, lemmata, tags

Analysis

- XSLT transformations to aggregate data for visualization software (SplitsTree, other)
- corpus queries and color-coded classification by XSLT based on the same *same configuration files*

Setup

Data format

- word aligned texts held in large XML files / CWB Corpus
- variables operationalized with regular expressions on word forms, lemmata, tags

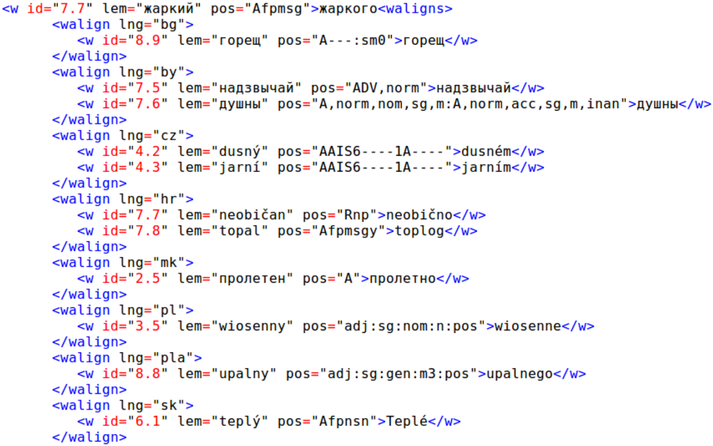
Analysis

- XSLT transformations to aggregate data for visualization software (SplitsTree, other)
- corpus queries and color-coded classification by XSLT based on the same *same configuration files*

to be published as open source in the near future

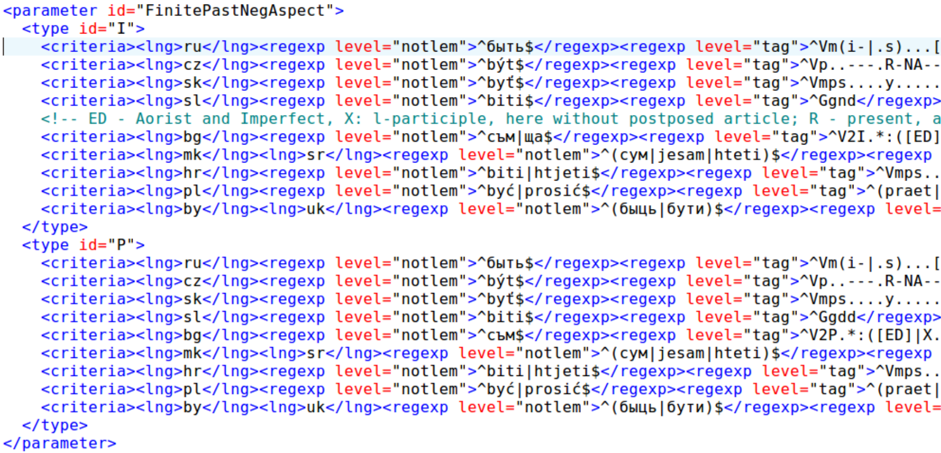
Word aligned corpus

- word aligned corpus held in large XML files



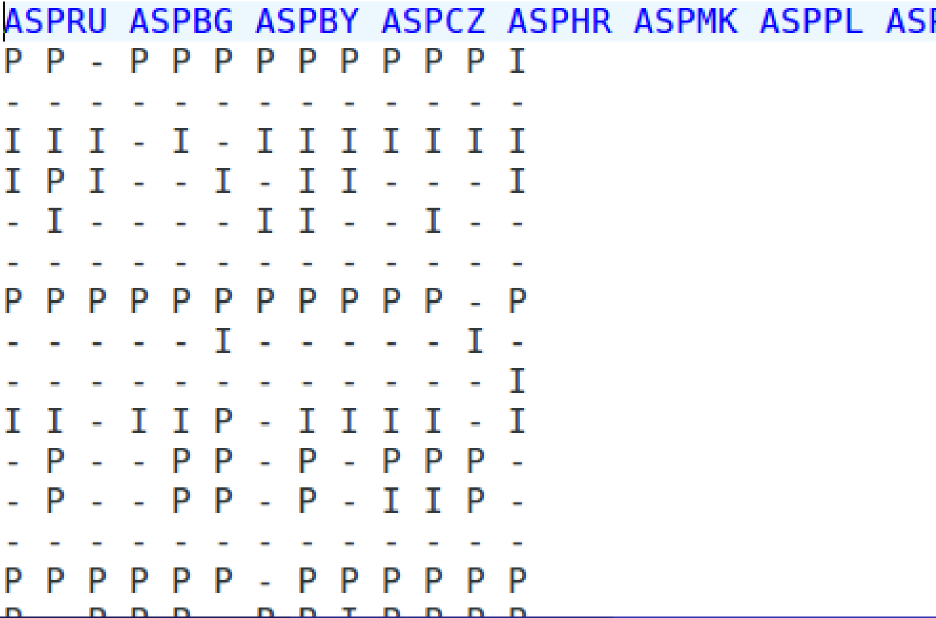
```
<w id="7.7" lem="жаркий" pos="Afpmsg">жаркого<waligns>
      <walign lng="bg">
         <w id="8.9" lem="горещ" pos="A---:sm0">горещ</w>
      </walign>
      <walign lng="by">
         <w id="7.5" lem="надзвычай" pos="ADV,norm">надзвычай</w>
         <w id="7.6" lem="душны" pos="A,norm,nom,sg,m:A,norm,acc,sg,m,inan">душны</w>
      </walign>
      <walign lng="cz">
         <w id="4.2" lem="dusný" pos="AAIS6----1A----">dusném</w>
         <w id="4.3" lem="jarní" pos="AAIS6----1A----">jarním</w>
      </walign>
      <walign lng="hr">
         <w id="7.7" lem="neobičan" pos="Rnp">neobično</w>
         <w id="7.8" lem="topal" pos="Afpmsgy">toplog</w>
      </walign>
      <walign lng="mk">
         <w id="2.5" lem="пролетен" pos="A">пролетно</w>
      </walign>
      <walign lng="pl">
         <w id="3.5" lem="wiosenny" pos="adj:sg:nom:n:pos">wiosenne</w>
      </walign>
      <walign lng="pla">
         <w id="8.8" lem="upalny" pos="adj:sg:gen:m3:pos">upalnego</w>
      </walign>
      <walign lng="sk">
         <w id="6.1" lem="teplý" pos="Afpnsn">Teplé</w>
      </walign>
```

Operationalization of variables



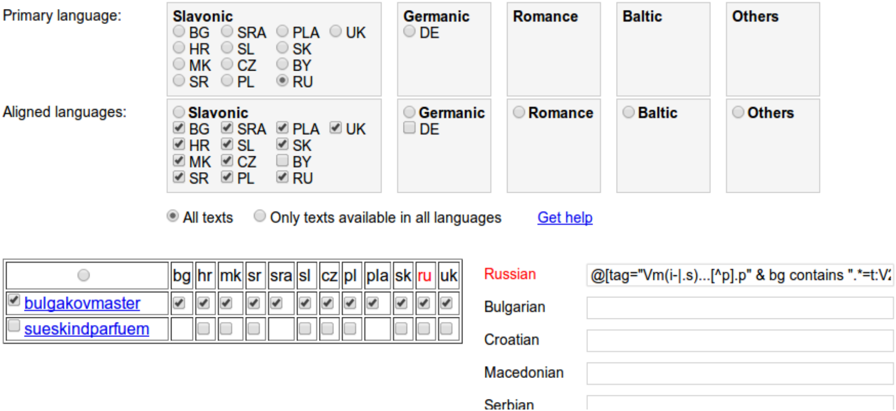
```
<parameter id="FinitePastNegAspect">
  <type id="I">
    <criteria><lng>ru</lng><regexp level="notlem">^быть$</regexp><regexp level="tag">^Vm(i-|.s)...[
    <criteria><lng>cz</lng><regexp level="notlem">^být$</regexp><regexp level="tag">^Vp..---.R-NA--
    <criteria><lng>sk</lng><regexp level="notlem">^byť$</regexp><regexp level="tag">^Vmps....y.....
    <criteria><lng>sl</lng><regexp level="notlem">^biti$</regexp><regexp level="tag">^Ggnd</regexp>
    <!-- ED - Aorist and Imperfect, X: l-participle, here without postposed article; R - present, a
    <criteria><lng>bg</lng><regexp level="notlem">^съм|ща$</regexp><regexp level="tag">^V2I.*:([ED]
    <criteria><lng>mk</lng><lng>sr</lng><regexp level="notlem">^(сум|jesam|hteti)$</regexp><regexp
    <criteria><lng>hr</lng><regexp level="notlem">^biti|htjeti$</regexp><regexp level="tag">^Vmps..
    <criteria><lng>pl</lng><regexp level="notlem">^być|prosić$</regexp><regexp level="tag">^(praet|
    <criteria><lng>by</lng><lng>uk</lng><regexp level="notlem">^(быць|бути)$</regexp><regexp level=
  </type>
  <type id="P">
    <criteria><lng>ru</lng><regexp level="notlem">^быть$</regexp><regexp level="tag">^Vm(i-|.s)...[
    <criteria><lng>cz</lng><regexp level="notlem">^být$</regexp><regexp level="tag">^Vp..---.R-NA--
    <criteria><lng>sk</lng><regexp level="notlem">^byť$</regexp><regexp level="tag">^Vmps....y.....
    <criteria><lng>sl</lng><regexp level="notlem">^biti$</regexp><regexp level="tag">^Ggdd</regexp>
    <criteria><lng>bg</lng><regexp level="notlem">^съм$</regexp><regexp level="tag">^V2P.*:([ED]|X.
    <criteria><lng>mk</lng><lng>sr</lng><regexp level="notlem">^(сум|jesam|hteti)$</regexp><regexp
    <criteria><lng>hr</lng><regexp level="notlem">^biti|htjeti$</regexp><regexp level="tag">^Vmps..
    <criteria><lng>pl</lng><regexp level="notlem">^być|prosić$</regexp><regexp level="tag">^(praet|
    <criteria><lng>by</lng><lng>uk</lng><regexp level="notlem">^(быць|бути)$</regexp><regexp level=
  </type>
</parameter>
```

Result table



```
ASPRU ASPBG ASPBY ASPCZ ASPHR ASPMK ASPPL ASP
P P - P P P P P P P P P I
- - - - - - - - - - - - -
I I I - I - I I I I I I I
I P I - - I - I I - - - I
- I - - - - I I - - I - -
- - - - - - - - - - - - -
P P P P P P P P P P P - P
- - - - - I - - - - - I -
- - - - - - - - - - - - I
I I - I I P - I I I I - I
- P - - P P - P - P P P -
- P - - P P - P - I I P -
- - - - - - - - - - - - -
P P P P P P - P P P P P P
P - P P P - P P I P P P P
```

Inspection of variable categorization in context



Primary language:

Slavonic
○ BG ○ SRA ○ PLA ○ UK
○ HR ○ SL ○ SK
○ MK ○ CZ ○ BY
○ SR ○ PL ● RU

Germanic
○ DE

Romance

Baltic

Others

Aligned languages:

○ **Slavonic**
☑ BG ☑ SRA ☑ PLA ☑ UK
☑ HR ☑ SL ☑ SK
☑ MK ☑ CZ □ BY
☑ SR ☑ PL ☑ RU

○ **Germanic**
□ DE

○ **Romance**

○ **Baltic**

○ **Others**

● All texts ○ Only texts available in all languages Get help

○	bg	hr	mk	sr	sra	sl	cz	pl	pla	sk	ru	uk
☑ bulgakovmaster	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
□ sueskindparfuem		□	□	□		□	□	□		□	□	□

Russian: @[tag="Vm(i-|.s)...[^p].p" & bg contains ".*=t:V

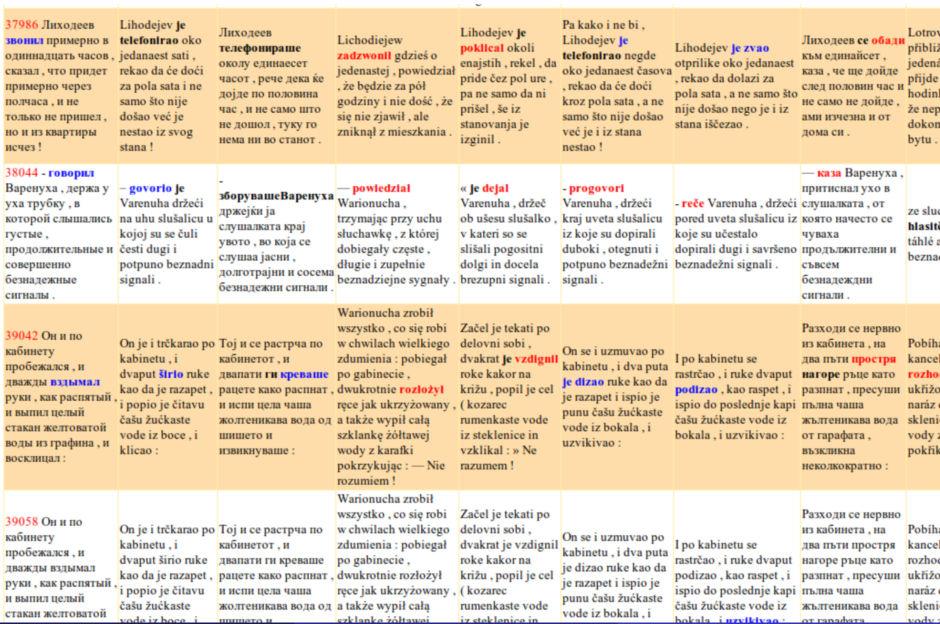
Bulgarian:

Croatian:

Macedonian:

Serbian

Inspection of variable categorization in context



37986 Лиходеев **звонил** примерно в одиннадцать часов , сказал , что придет примерно через полчаса , и не только не пришел , но и из квартиры исчез !	Lihodejev **je telefonirao** oko jedanaest sati , rekao da će doći za pola sata i ne samo što nije došao već je nestao iz svog stana !	Лиходеев **телефонираше** околу единаесет часот , рече дека ќе дојде по половина час , и не само што не дошол , туку го нема ни во станот .	Lichodiejew **zadzwonił** gdzieś o jedenastej , powiedział , że będzie za pół godziny i nie dość , że się nie zjawił , ale zniknął z mieszkania .	Lihodejev **je poklical** okoli enajstih , rekel , da pride čez pol ure , pa ne samo da ni prišel , še iz stanovanja je izginil .	Pa kako i ne bi , Lihodejev **je telefonirao** negde oko jedanaest časova , rekao da će doći kroz pola sata , a ne samo što nije došao već je i iz stana nestao !	Lihodejev **je zvao** otprilike oko jedanaest , rekao da dolazi za pola sata , a ne samo što nije došao nego je i iz stana iščezao .	Лиходеев **се обади** към единайсет , каза , че ще дойде след половин час и не само не дойде , ами изчезна и от дома си .	Lotrov přibliž jedená přijde hodink že nep dokon bytu .
38044 - **говорил** Варенуха , держа у уха трубку , в которой слышались густые , продолжительные и совершенно безнадежные сигналы .	– **govorio je** Varenuha držeći na uhu slušalicu u kojoj su se čuli česti dugi i potpuno beznadni signali .	- **зборувашеВаренуха** држејќи ја слушалката крај увото , во која се слушаа јасни , долготрајни и сосема безнадежни сигнали .	— **powiedział** Warionucha , trzymając przy uchu słuchawkę , z której dobiegały częste , długie i zupełnie beznadziejne sygnały .	« **je dejal** Varenuha , držeč ob ušesu slušalko , v kateri so se slišali pogositni dolgi in docela brezupni signali .	- **progovori** Varenuha , držeći kraj uveta slušalicu iz koje su dopirali duboki , otegnuti i potpuno beznadežni signali .	- **reče** Varenuha , držeći pored uveta slušalicu iz koje su učestalo dopirali dugi i savršeno beznadežni signali .	— **каза** Варенуха , притиснал ухо в слушалката , от която начесто се чуваха продължителни и съвсем безнадеждни сигнали .	ze sluc **hlasitě** táhlé a beznad
39042 Он и по кабинету пробежался , и дважды **вздымал** руки , как распятый , и выпил целый стакан желтоватой воды из графина , и восклицал :	On je i trčkarao po kabinetu , i dvaput **širio** ruke kao da je razapet , i popio je čitavu čašu žućkaste vode iz boce , i klicao :	Тој и се растрча по кабинетот , и двапати **ги креваше** рацете како распнат , и испи цела чаша жолтеникава вода од шишето и извикнуваше :	Warionucha zrobił wszystko , co się robi w chwilach wielkiego zdumienia : pobiegał po gabinecie , dwukrotnie **rozłożył** ręce jak ukrzyżowany , a także wypił całą szklankę żółtawej wody z karafki pokrzykując : — Nie rozumiem !	Začel je tekati po delovni sobi , dvakrat **je vzdignil** roke kakor na križu , popil je cel (kozarec rumenkaste vode iz steklenice in vzklikal : » Ne razumem !	On se i uzmuvao po kabinetu , i dva puta **je dizao** ruke kao da je razapet i ispio je punu čašu žućkaste vode iz bokala , i uzvikivao :	I po kabinetu se rastrčao , i ruke dvaput **podizao** , kao raspet , i ispio do poslednje kapi čašu žućkaste vode iz bokala , i uzvikivao :	Разходи се нервно из кабинета , на два пъти **простря нагоре** ръце като разпнат , пресуши пълна чаша жълтеникава вода от гарафата , възкликна неколкократно :	Pobíha kancel **rozhod** ukřižo naráz sklení vody z pokřik
39058 Он и по кабинету пробежался , и дважды вздымал руки , как распятый , и выпил целый стакан желтоватой	On je i trčkarao po kabinetu , i dvaput širio ruke kao da je razapet , i popio je čitavu čašu žućkaste vode iz boce , i	Тој и се растрча по кабинетот , и двапати ги креваше рацете како распнат , и испи цела чаша жолтеникава вода од шишето и	Warionucha zrobił wszystko , co się robi w chwilach wielkiego zdumienia : pobiegał po gabinecie , dwukrotnie rozłożył ręce jak ukrzyżowany , a także wypił całą szklankę żółtawej	Začel je tekati po delovni sobi , dvakrat je vzdignil roke kakor na križu , popil je cel (kozarec rumenkaste vode iz steklenice in	On se i uzmuvao po kabinetu , i dva puta je dizao ruke kao da je razapet i ispio je punu čašu žućkaste vode iz bokala , i	I po kabinetu se rastrčao , i ruke dvaput podizao , kao raspet , i ispio do poslednje kapi čašu žućkaste vode iz bokala , i **uzvikivao** :	Разходи се нервно из кабинета , на два пъти простря нагоре ръце като разпнат , пресуши пълна чаша жълтеникава вода от гарафата	Pobíha kancel rozhod ukřižo naráz sklení vody z

Automatization: summary

- operationalization of variables: regexps on token/lemma/tag level with word aligned files
- error control and fine-tuning of operationalization
  - define variables
  - inspect aggregation level,
  - inspect level of corpus examples
  - redefine variables

→ allows fast aggregation of *different* variables

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Compare the use of verbal prefixes across Slavic

(7) Lovite že ego nemedlenno, ...
‘Catch him immediately,...’

a. [RU] inače on *natvorit* neopisuemyx bed !
b. [UK] a to vin *nakoit’* lixa - bidi !
c. [PL] inaczej *narobi* strasznych rzeczy!
d. [PL] inaczej *narozrabia* tak , że się nie pozbieramy !
e. [CZ] nebo *natropí* nepopsatelné škody !
f. [SK] lebo *narobí* ukrutne veľa zla !
g. [SL] sicer bo *naredil* nepopisno škodo !
h. [HR] inače će *počiniti* neopisive nevolje !
i. [SR] ili će svekolike nedaće *počiniti* !
j. [SR] inače će *učiniti* neopisivo zlo !
k. [MK] inaku ḱe *sozdade* nevideni nevolji !
l. [BG] inače toj šte *izvărši* neopisuemi porazii !
‘ ... otherwise he’ll *do* untold harm!’

Operationalization

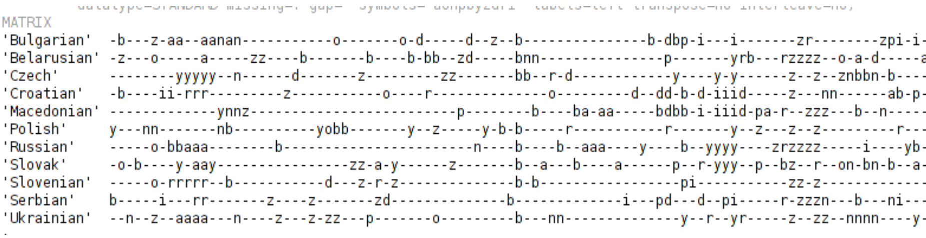
Classify each verb form as belonging to one of eleven prefix classes (*do*-, *pered/pred/pro*-, *vy*-, *za*-, *na*-,*u*-, *od*-, *iz*-, *vy*-, *raz*-, *po*-); operationalize with regular expressions on lemmata / tags

	NA	OT	IZ	VY	PERE	RAZ	PRI	PO	DO	ZA
ru	^na	^ot	^i[zs]	^vy	^pere	^ra[zs]	^pri	^po[^d]	^do	^za
bg	^na	^ot	^iz	^vi	^pre	^raz	^pri	^po[^d]	^do	^za
by	^na	^ad	^[ii]z	^vy	^pera	^ra[zs]	^pry	^pa[^d]	^da	^za
cz	^na	^od	^iz	^vy	^pře	^raz	^při	^po[^d]	^do	^za
hr	^na	^o[dt]	^i[zs]	^vi	^pre	^raz	^pri	^po[^d]	^do	^za
mk	^na	^o[dt]	^i[zs]	^vi	^pre	^raz	^pri	^po[^d]	^do	^za
pl	^na	^od	^iz	^wy	^prze	^roz	^przy	^po[^d]	^do	^za
sk	^na	^od	^iz	^vy	^pre	^raz	^pri	^po[^d]	^do	^za
sl	^na	^od	^iz	^vi	^pre	^raz	^pri	^po[^d]	^do	^za
sr	^na	^o[dt]	^i[zs]	^vi	^pre	^raz	^pri	^po[^d]	^do	^za
uk	^na	^v[ii]d	^[ii]z	^vi	^pere	^ro[zs]	^pri	^po[^d]	^do	^za

Table : List of verbal prefix types and how they are classified (transliterated)

Operationalization

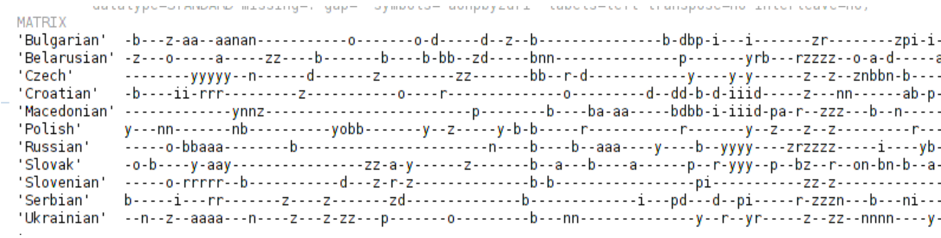
Use word alignment to derive a matrix of values



```
        datatype=STANDARD missing=? gap=- symbols= aonpbyzdri  labels=left transpose=no interleave=no;
MATRIX
'Bulgarian'  -b---z-aa--aanan-----------o-------o-d-----d--z--b--------------b-dbp-i---i-------zr--------zpi-i-
'Belarusian' -z---o-----a-----zz----b-------b----b-bb--zd-----bnn--------------p-------yrb---rzzzz--o-a-d-----a
'Czech'      --------yyyyy--n------d-------z---------zz-------bb--r-d------------y----y-y------z--z--znbbn-b----
'Croatian'   -b----ii-rrr---------z------------o----r--------------o---------d--dd-b-d-iiid-----z---nn------ab-p-
'Macedonian' -------------ynnz-------------------------p--------b----ba-aa-----bdbb-i-iiid-pa-r--zzz---b--n-----
'Polish'     y---nn-------nb-----------yobb-------y--z-----y-b-b-----r-----------r-------y--z---z--z---------r---
'Russian'    -----o-bbaaa--------b-------------------------n----b----b--aaa----y----b--yyyy---zrzzzz-----i----yb-
'Slovak'     -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p--r-yyy--p--bz--r--on-bn-b--a-
'Slovenian'  -----o-rrrrr--b-----------d---z-r-z-------------b-b-----------------pi-----------zz-z------------
'Serbian'    b-----i---rr-------z----z-------zd--------------b-------------i---pd---d--pi-----r-zzzn---b---ni---
'Ukrainian'  --n--z--aaaa---n----z---z-zz---p-------o---------b---nn---------------y--r--yr-----z--zz--nnnn----y-
;
```

Operationalization

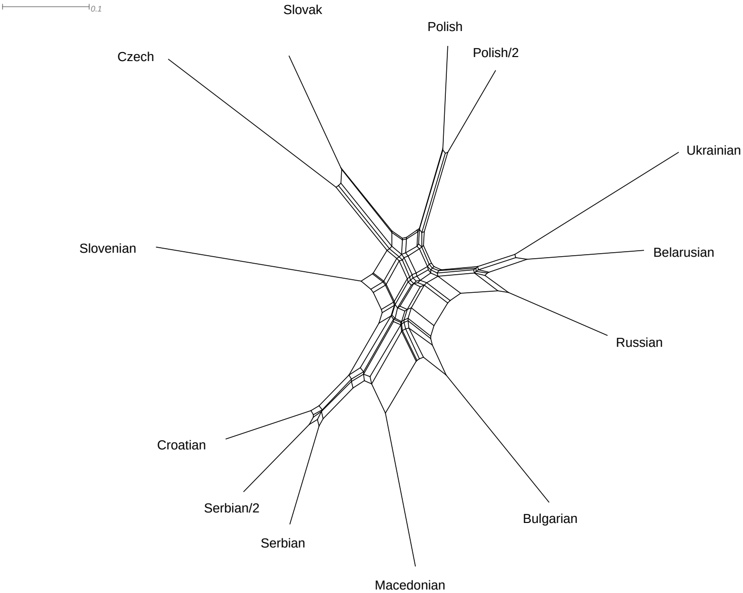
Use word alignment to derive a matrix of values



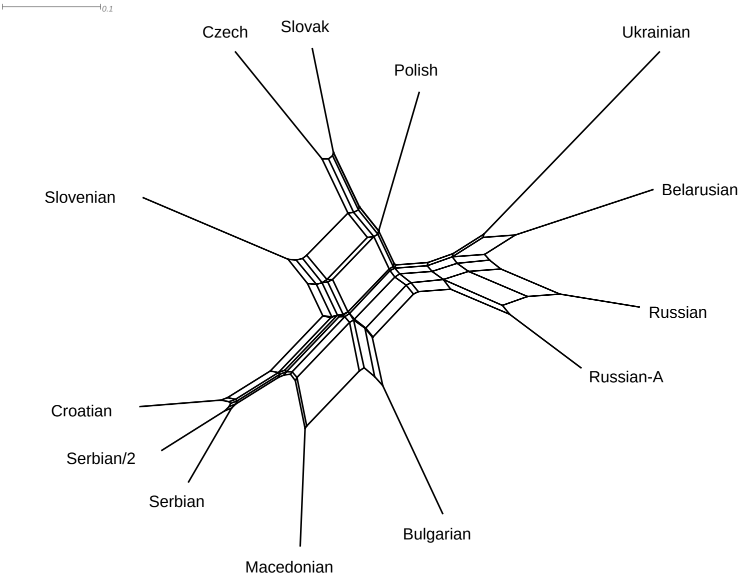
```
        datatype=STANDARD missing=? gap=- symbols="aonpbyzdri" labels=left transpose=no interleave=no;
MATRIX
'Bulgarian'  -b---z-aa--aanan-----------o-------o-d-----d--z--b---------------b-dbp-i---i-------zr--------zpi-i-
'Belarusian' -z---o-----a-----zz----b-------b----b-bb--zd-----bnn---------------p-------yrb---rzzzz--o-a-d-----a
'Czech'      --------yyyyy--n------d-------z---------zz-------bb--r-d------------y----y-y------z--z--znbbn-b----
'Croatian'   -b----ii-rrr---------z-----------o----r--------------o---------d--dd-b-d-iiid-----z---nn------ab-p-
'Macedonian' -------------ynnz-------------------------p--------b----ba-aa-----bdbb-i-iiid-pa-r--zzz---b--n-----
'Polish'     y---nn-------nb----------yobb-------y--z-----y-b-b-----r-----------r-------y--z---z--z---------r---
'Russian'    -----o-bbaaa--------b-----------------------n----b----b--aaa----y----b--yyyy----zrzzzz-----i----yb-
'Slovak'     -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p--r-yyy--p--bz--r--on-bn-b--a-
'Slovenian'  -----o-rrrrr--b-----------d---z-r-z--------------b-b-----------------pi-----------zz-z-------------
'Serbian'    b-----i---rr-------z----z-------zd--------------b-------------i---pd---d--pi-----r-zzzn---b---ni---
'Ukrainian'  --n--z--aaaa---n----z---z-zz---p-------o---------b---nn--------------y--r--yr-----z--zz--nnnn----y-
;
```

… visualize in SplitsTree

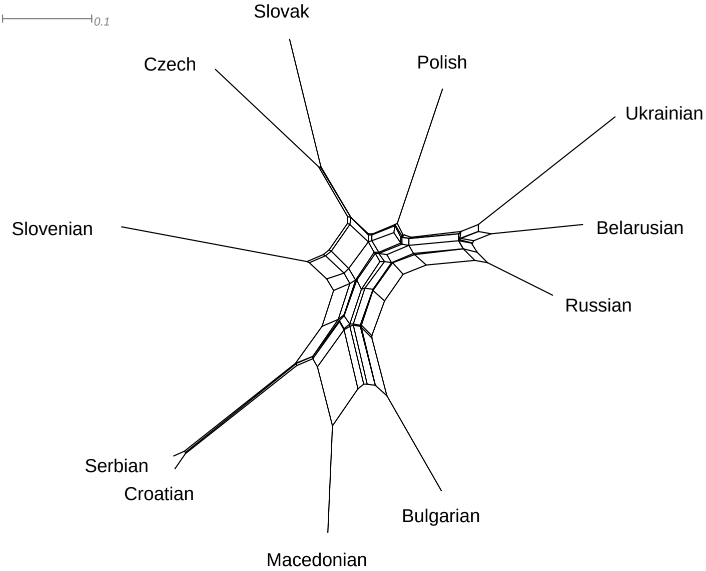
Bulgakov, verbal prefixes



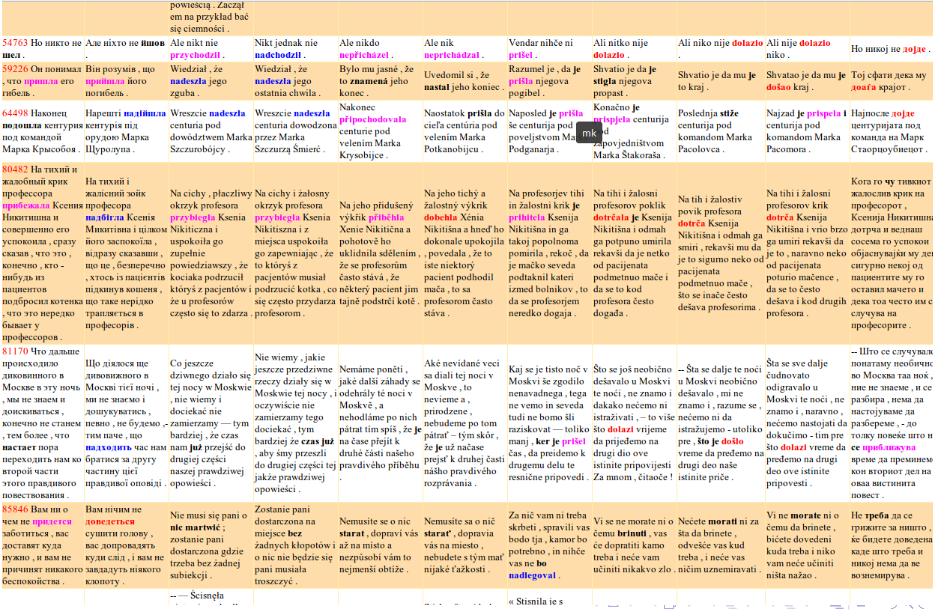
Solaris, verbal prefixes



Ostrovski, verbal prefixes



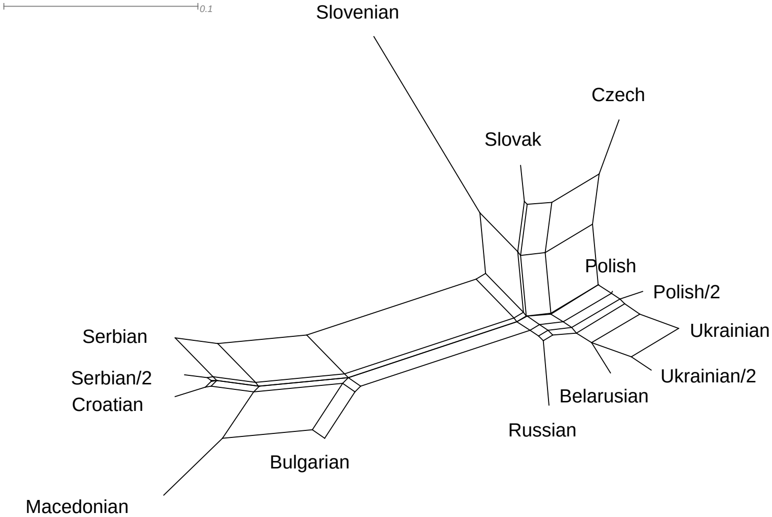
Inspect data points in browser



		powieścią . Zacząłem na przykład bać się ciemności .								
54763 Но никто не **шел** .	Але ніхто не **йшов** .	Ale nikt nie **przychodził** .	Nikt jednak nie **nadchodził** .	Ale nikdo **nepřicházel** .	Ale nik **neprichádzal** .	Vendar nihče ni **prišel** .	Ali nitko nije **dolazio** .	Ali niko nije **dolazio** .	Ali nije **dolazio** niko .	Но никој не **дојде** .
59226 Он понимал , что **пришла** его гибель .	Він розумів , що **прийшла** його погибель .	Wiedział , że **nadeszła** jego zguba .	Wiedział , że **nadeszła** jego ostatnia chwila .	Bylo mu jasné , že to **znamená** jeho konec .	Uvedomil si , že **nastal** jeho koniec .	Razumel je , da **je prišla** njegova pogibel .	Shvatio je da **je stigla** njegova propast .	Shvatio je da mu **je** to kraj .	Shvatao je da mu **je došao** kraj .	Тој сфати дека му **доаѓа** крајот .
64498 Наконец **подошла** кентурия под командой Марка Крысобоя .	Нарешті **надійшла** кентурія під орудою Марка Щуролупа .	Wreszcie **nadeszła** centuria pod dowództwem Marka Szczurobójcy .	Wreszcie **nadeszła** centuria dowodzona przez Marka Szczurzą Śmierć .	Nakonec **připochodovala** centurie pod velením Marka Krysobijce .	Naostatok **prišla** do cieľa centúria pod velením Marka Potkanobijcu .	Naposled **je prišla** še centurija pod poveljstvom Marka Podganarja .	Konačno **je prispjela** centurija pod zapovjedništvom Marka Štakoraša .	Poslednja **stiže** centurija pod komandom Marka Pacolovca .	Najzad **je prispela i** centurija pod komandom Marka Pacomora .	Најпосле **дојде** центуријата под команда на Марк Стаорцоубиецот .
80482 На тихий и жалобный крик профессора **прибежала** Ксения Никитишна и совершенно его успокоила , сразу сказав , что это , конечно , кто - нибудь из пациентов подбросил котенка , что это нередко бывает у профессоров .	На тихий і жалісний зойк професора **надбігла** Ксенія Микитівна і цілком його заспокоїла , відразу сказавши , що це , безперечно , хтось із пацієнтів підкинув кошеня , що таке нерідко трапляється в професорів .	Na cichy , płaczliwy okrzyk profesora **przybiegła** Ksenia Nikiticzna i uspokoiła go zupełnie powiedziawszy , że kociaka podrzucił któryś z pacjentów i że u profesorów często się to zdarza .	Na cichy i żałosny okrzyk profesora **przybiegła** Ksenia Nikitiszna i z miejsca uspokoiła go zapewniając , że to któryś z pacjentów musiał podrzucić kotka , co się często przydarza profesorom .	Na jeho přidušený výkřik **přiběhla** Xenie Nikitična a pohotově ho uklidnila sdělením , že se profesorům často stává , že některý pacient jim tajně podstrčí kotě .	Na jeho tichý a žalostný výkrik **dobehla** Xénia Nikitišna a hneď ho dokonale upokojila , povedala , že to iste niektorý pacient podhodil mača , to sa profesorom často stáva .	Na profesorjev tihi in žalostni krik **je prihitela** Ksenija Nikitišna in ga takoj popolnoma pomirila , rekoč , da je mačko seveda podtaknil kateri izmed bolnikov , to da se profesorjem neredko dogaja .	Na tihi i žalosni profesorov poklik **dotrčala je** Ksenija Nikitišna i odmah ga potpuno umirila rekavši da je netko od pacijenata podmetnuo mače i da se to kod profesora često događa .	Na tih i žalostiv povik profesora **dotrča** Ksenija Nikitišna i odmah ga smiri , rekavši mu da je to sigurno neko od pacijenata podmetnuo mače , što se inače često dešava profesorima .	Na tihi i žalosni profesorov krik **dotrča** Ksenija Nikitišna i vrio brzo ga umiri rekavši da je to , naravno neko od pacijenata poturio mačence , da se to često dešava i kod drugih profesora .	Кога го **чу** тивкиот жалослив крик на професорот , Ксенија Никитишна дотрча и веднаш сосема го успокои објаснувајќи му де сигурно некој од пациентите му го оставил мачето и дека тоа често им с случува на професорите .
81170 Что дальше происходило диковинного в Москве в эту ночь , мы не знаем и доискиваться , конечно не станем , тем более , что **настает** пора переходить нам ко второй части этого правдивого повествования .	Що діялося ще дивовижного в Москві тієї ночі , ми не знаємо і дошукуватись , певно , не будемо ,- тим паче , що **надходить** час нам братися за другу частину цієї правдивої оповіді .	Co jeszcze dziwnego działo się tej nocy w Moskwie , nie wiemy i dociekać nie zamierzamy — tym bardziej , że czas nam **już** przejść do drugiej części naszej prawdziwej opowieści .	Nie wiemy , jakie jeszcze przedziwne rzeczy działy się w Moskwie tej nocy , i oczywiście nie zamierzamy tego dociekać , tym bardziej że **czas już** , aby śmy przeszli do drugiej części tej jakże prawdziwej opowieści .	Nemáme ponětí , jaké další záhady se odehrály té noci v Moskvě , a nehodláme po nich pátrat tím spíš , že **je** na čase přejít k druhé části našeho pravdivého příběhu .	Aké nevídané veci sa diali tej noci v Moskve , to nevieme a , prirodzene , nebudeme po tom pátrať – tým skôr , že **je** už načase prejsť k druhej časti nášho pravdivého rozprávania .	Kaj se je tisto noč v Moskvi še zgodilo nenavadnega , tega ne vemo in seveda tudi ne bomo šli raziskovat — toliko manj , **ker je prišel** čas , da preidemo k drugemu delu te resnične pripovedi .	Što se još neobično dešavalo u Moskvi te noći , ne znamo i dakako nećemo ni istraživati , – to više što **dolazi** vrijeme da prijeđemo na drugi dio ove istinite pripovijesti Za mnom , čitaoče !	-- Šta se dalje te noći u Moskvi neobično dešavalo , mi ne znamo i , razume se , nećemo ni da istražujemo - utoliko pre , **što je došlo** vreme da pređemo na drugi deo naše istinite priče .	Šta se sve dalje čudnovato odigravalo u Moskvi te noći , ne znamo i , naravno , nećemo nastojati da dokučimo - tim pre što **dolazi** vreme da pređemo na drugi deo ove istinite pripovesti .	-- Што се случувал понатаму необично во Москва таа ноќ , ние не знаеме , и се разбира , нема да настојуваме да разбереме , - до толку повеќе што н **се приближува** време да преминем кон вториот дел на оваа вистинита повест .
85846 Вам ни о чем не **придется** заботиться , вас доставят куда нужно , и вам не причинят никакого беспокойства .	Вам нічим не **доведеться** сушити голову , вас допровадять куди слід , і вам не завдадуть ніякого клопоту .	Nie musi się pani o **nic martwić** ; zostanie pani dostarczona gdzie trzeba bez żadnej subiekcji .	Zostanie pani dostarczona na miejsce **bez** żadnych kłopotów i o nic nie będzie się pani musiała troszczyć .	Nemusíte se o nic **starat** , dopraví vás až na místo a nezpůsobí vám to nejmenší obtíže .	Nemusíte sa o nič **starať** , dopravia vás na miesto , nebudete s tým mať nijaké ťažkosti .	Za nič vam ni treba skrbeti , spravili vas bodo tja , kamor bo potrebno , in nihče vas ne **bo nadlegoval** .	Vi se ne morate ni o čemu **brinuti** , vas će dopratiti kamo treba i neće vam učiniti nikakvo zlo .	Nećete **morati** ni za šta da brinete , odvešće vas kud treba , i neće vas ničim uznemiravati .	Vi ne **morate** ni o čemu da brinete , bićete dovedeni kuda treba i niko vam neće učiniti ništa nažao .	Не **треба** да се грижите за ништо , ќе бидете доведена каде што треба и никој нема да ве вознемирува .
		-- — Ścisnęła			[illegible]	« Stisnila je s				

Comparing individual prefiexes

Use of *pri-* vs. *do-* in Bulgakov



Conclusions

- interesting results concerning crosslinguistic use of cognate derivational affixes
- good *genealogical* signal, different from aspect

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Compare the use of nominal suffixes across Slavic

(8) a. [RU] Vot kogo s osobennym udovol'stviem otpušču, - skazal Voland, s otvraščeniem gljadja na Nikolaja Ivanoviča, - s iskljucitel'nym udovol'stviem, nastol'ko on zdes' lišnij.

b. [UK] Os' kogo z osoblivim zadovolennjam vidpušču

c. [PL] Tego odprawię ze szczególną przyjemnością

d. [CZ] Toho propustím se zvláštním potěšením,

e. [SK] Tohto prepustím s radosťou,

f. [SL] Tu je nekdo, ki ga odpuščam s posebnim zadovoljstvom

g. [HR] Evo, koga ću sa zadovoljstvom otpustiti

h. [SR] A ovoga ću sa velikim zadovoljstvom pustiti

i. [MK] Ete kogo ḱe go puštam so osobeno zadovolstvo

j. [BG] Viž, nego s naj goljamo udovolstvie šče go pusna da si v'rvi

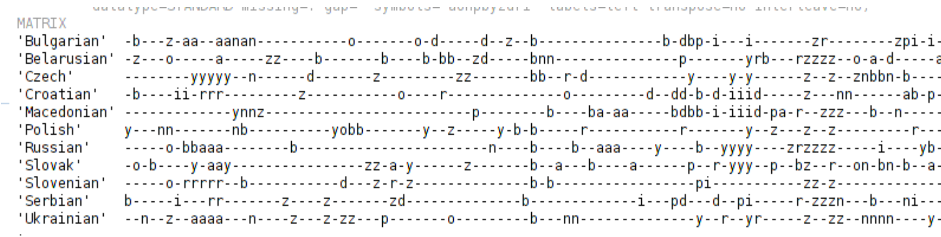
Operationalization

Classify each verb form as belonging to a suffix class, operationalized with regular expressions on lemmata / tags

	ICA	TEL	NIK	AR	AK	IK	EC	OST	STVO	STVIE	NIE
ru	[iy]ca$	tel’$	nik$	ar’$	[jaa]k$	[^n]ik$	ec$	ost’$	stvo$	stvie$	n[’i]e$
uk	icja$	tel’$	nik$	ar$	[jaa]k$$	[^n]ik$	ec’$	ist’$	[sc]tvo$	XXX$	nnja$
by	[iy]ca$	cel’$	nik$	ar$	[jaa]k$$	[^n]ik$	ec$	asc’$	[sc]tva$	XXX$	nne$
mk	ica$	tel$	nik$	ar$	ak$	[^n]ik$	ec$	ost$	[šs]tvo$	XXX$	nje$
bg	ica$	tel$	nik$	ar$	[jaa]k$	[^n]ik$	ec$	ost$	stvo$	XXX$	nie$
pl	[iy]ca$	ciel$	nik$	arz$	ak$	[^n]ik$	ec$	ość$	[cs]two$	XXX$	nie$
cz	ice$	tel$	n[ií]k$	[áa]r$	[aá]k$	[^n][ií]k$	ec$	ost$	stvo$	ství$	ní$
sk	ica$	teľ$	n[ií]k$	[áa]r$	[aá]k$	[^n][ií]k$	ec$	osť$	stvo$	XXX$	nie$
sl	ica$	telj$	nik$	ar$	ak$	[^n]ik$	ec$	ost$	[šs]tvo$	XXX$	nje$
hr	ica$	telj$	nik$	ar$	ak$	[^n]ik$	ac$	ost$	[šs]tvo$	XXX$	nje$
sr	ica$	telj$	nik$	ar$	ak$	[^n]ik$	ac$	ost$	[šs]tvo$	XXX$	nje$

Operationalization

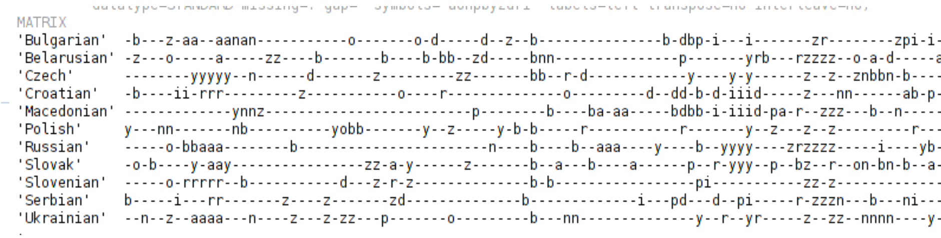
Use word alignment to derive a matrix of values



```
         datatype=STANDARD missing=? gap=- symbols="aonpbyzdri" labels=left transpose=no interleave=no;
  MATRIX
  'Bulgarian'  -b---z-aa--aanan-----------o-------o-d-----d--z--b---------------b-dbp-i---i-------zr--------zpi-i-
  'Belarusian' -z---o-----a-----zz----b-------b----b-bb--zd-----bnn---------------p-------yrb---rzzzz--o-a-d-----a
  'Czech'      --------yyyyy--n------d-------z---------zz-------bb--r-d------------y----y-y------z--z--znbbn-b----
  'Croatian'   -b----ii-rrr---------z-----------o----r--------------o---------d--dd-b-d-iiid-----z---nn------ab-p-
  'Macedonian' -------------ynnz-------------------------p--------b----ba-aa-----bdbb-i-iiid-pa-r--zzz---b--n-----
  'Polish'     y---nn-------nb----------yobb-------y--z-----y-b-b-----r-----------r-------y--z---z--z---------r---
  'Russian'    -----o-bbaaa--------b-----------------------n----b----b--aaa----y----b--yyyy----zrzzzz-----i----yb-
  'Slovak'     -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p--r-yyy--p--bz--r--on-bn-b--a-
  'Slovenian'  -----o-rrrrr--b-----------d---z-r-z--------------b-b-----------------pi-----------zz-z-------------
  'Serbian'    b-----i---rr-------z----z-------zd--------------b-------------i---pd---d--pi-----r-zzzn---b---ni---
  'Ukrainian'  --n--z--aaaa---n----z---z-zz---p-------o---------b---nn--------------y--r--yr-----z--zz--nnnn----y-
  ;
```

Operationalization

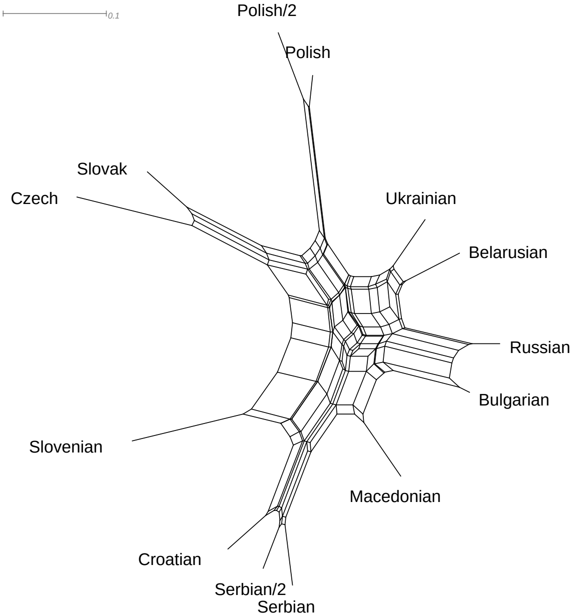
Use word alignment to derive a matrix of values



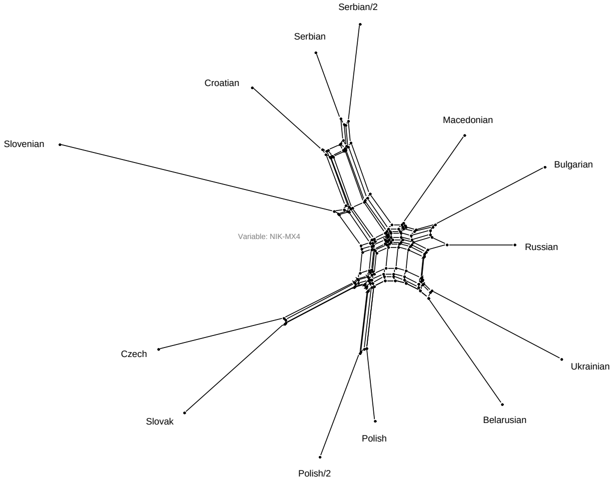
```
        datatype=STANDARD missing=? gap=- symbols="aonpbyzdri" labels=left transpose=no interleave=no;
MATRIX
'Bulgarian'  -b---z-aa--aanan-----------o-------o-d-----d--z--b--------------b-dbp-i---i-------zr--------zpi-i-
'Belarusian' -z---o-----a-----zz----b-------b----b-bb--zd-----bnn---------------p-------yrb---rzzzz--o-a-d-----a
'Czech'      --------yyyyy--n------d-------z---------zz-------bb--r-d------------y----y-y------z--z--znbbn-b----
'Croatian'   -b----ii-rrr---------z-----------o----r--------------o---------d--dd-b-d-iiid-----z---nn-------ab-p-
'Macedonian' -------------ynnz-------------------------p--------b----ba-aa-----bdbb-i-iiid-pa-r--zzz---b--n-----
'Polish'     y---nn-------nb----------yobb-------y--z-----y-b-b-----r------------r-------y--z---z--z---------r---
'Russian'    -----o-bbaaa--------b------------------------n----b----b--aaa----y----b--yyyy----zrzzzz-----i----yb-
'Slovak'     -o-b----y-aay----------------zz-a-y------z-------b--a---b----a------p--r-yyy--p--bz--r--on-bn-b--a-
'Slovenian'  -----o-rrrrr--b-----------d---z-r-z--------------b-b------------------pi-----------zz-z-------------
'Serbian'    b-----i---rr-------z----z-------zd-------------b--------------i---pd---d--pi-----r-zzzn---b---ni---
'Ukrainian'  --n--z--aaaa---n----z---z-zz---p-------o---------b---nn---------------y--r--yr-----z--zz--nnnn----y-
.
```

... visualize in SplitsTree

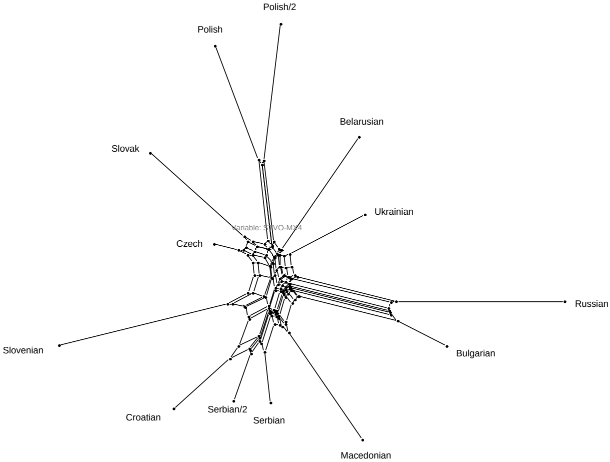
Bulgakov, nominal suffixes



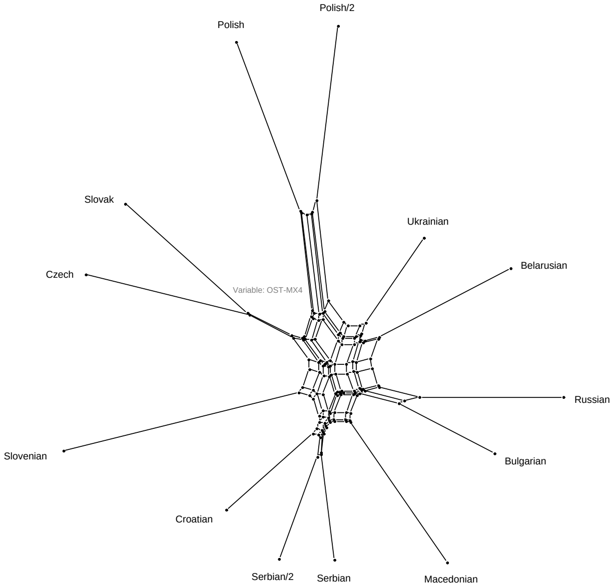
Bulgakov, -NIK



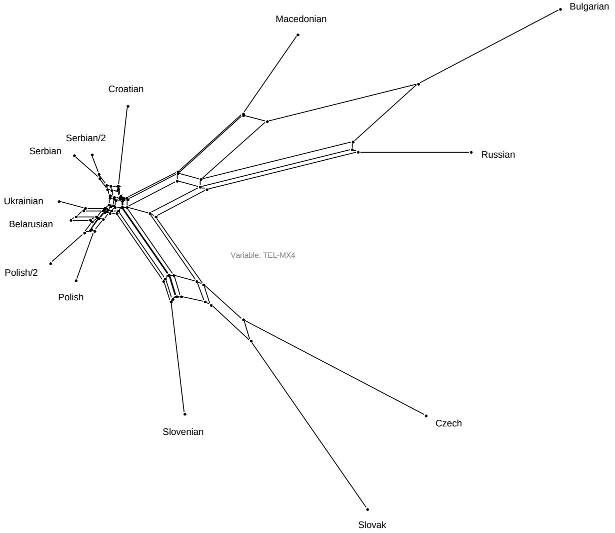
Bulgakov, -STVO



Bulgakov, -OST



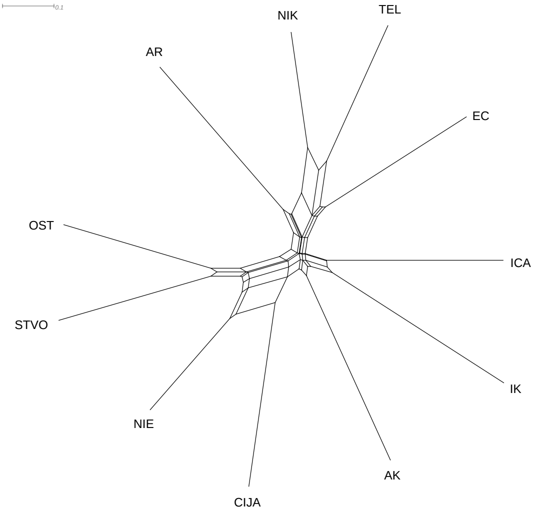
Bulgakov, -TELJ



Reverse perspective: clustering suffixes

How similar are the suffixes in their functional domains, i.e., contexts?

Reverse perspective: clustering suffixes

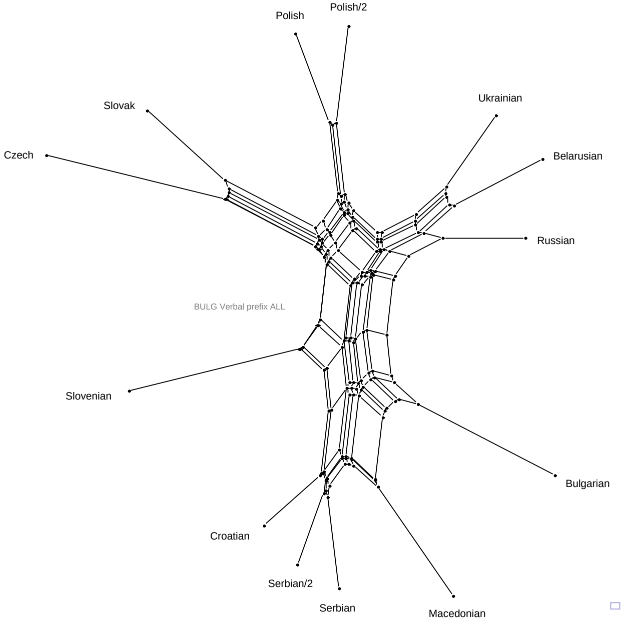


Pairwise distances: intersection (suff1, suff2) / max (suff1suff2)

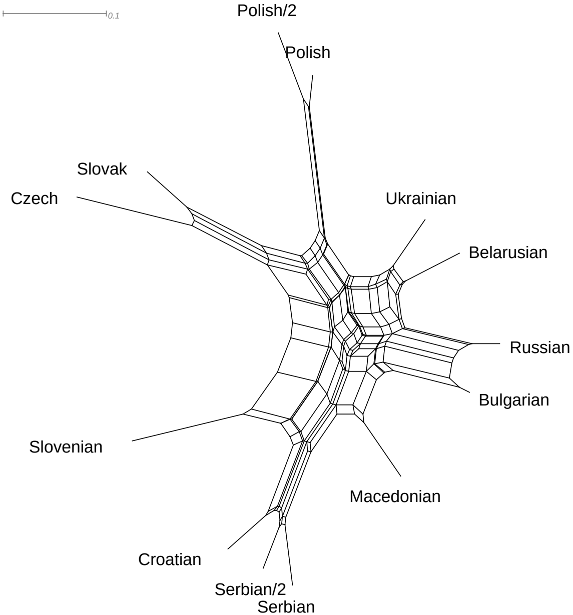
Comparing verbal prefixes and nominal suffixes

Differences between nominal and verbal domain?

Bulgakov, verbal prefixes



Bulgakov, nominal suffixes



Conclusions

- A lot of contact based influence with nominal suffixes
- Much less so with verbal prefixes
- In line with general knowledge (see Haspelmath 2009) - but rules out explanation of differences in nominal vs. verbal morphological complexity

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Participle vs. finite subordinate clause

Eg. Russian original and Serbian translation:

(9)

Popav	v	ten'	čut'	**zelenejuščix**
arrived.PST.CVB	in	shadow	a-bit	green.PRES.PCPL.GEN.PL

lip,	pisateli	[brosilis'	k	budočke].
linden.GEN.PL.F	writers.INF.PL	dashed	to	stand.DAT.SG.F

(10)

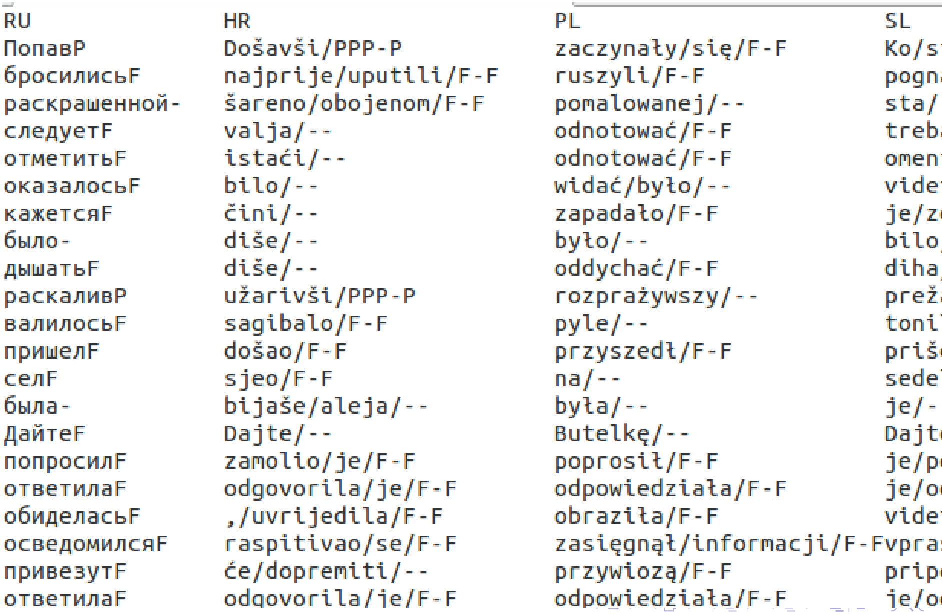
Čim	**su**	**dospeli**	u	senku	lipa,
When	3PL	arrive.PST.PL	in	shadow.LOC	linden.GEN.PL

koje	su	tek	**počinjale**	**da**
REL.NOM.PL	3PL	just	start.PST.PL	COMP

zelene,	pisci	najpre	pohitaše...
become-green.3PL	authors-NOM.PL	first	hurried...

'Once in the shade of the barely greening lindens, the writers dashed first thing to a brightly painted stand with the sign: 'Beer and Soft Drinks.' (Bulgakov)

Word alignment

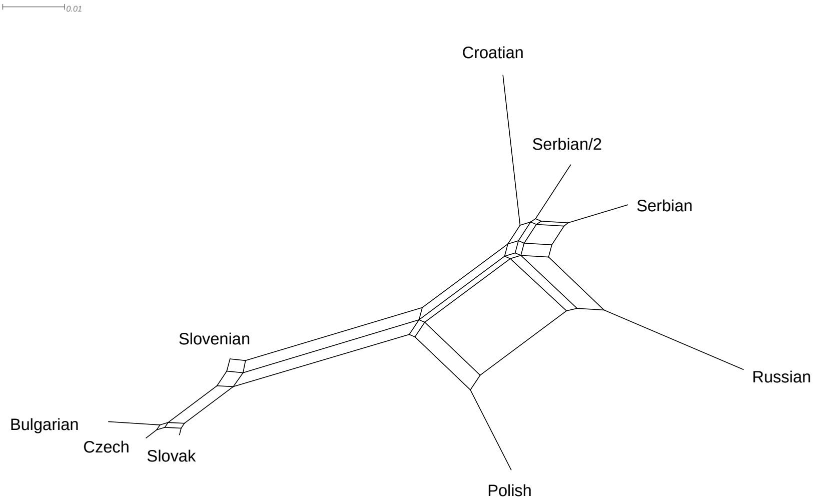


RU	HR	PL	SL
ПопавP	Došavši/PPP-P	zaczynały/się/F-F	Ko/s
бросилисьF	najprije/uputili/F-F	ruszyli/F-F	pogn
раскрашенной-	šareno/obojenom/F-F	pomalowanej/--	sta/
следуетF	valja/--	odnotować/F-F	treb
отметитьF	istaći/--	odnotować/F-F	omen
оказалосьF	bilo/--	widać/było/--	vide
кажетсяF	čini/--	zapadało/F-F	je/z
было-	diše/--	było/--	bilo
дышатьF	diše/--	oddychać/F-F	diha
раскаливP	užarivši/PPP-P	rozprażywszy/--	prež
валилосьF	sagibalo/F-F	pyle/--	toni
пришелF	došao/F-F	przyszedł/F-F	priš
селF	sjeo/F-F	na/--	sede
была-	bijaše/aleja/--	była/--	je/-
ДайтеF	Dajte/--	Butelkę/--	Dajt
попросилF	zamolio/je/F-F	poprosił/F-F	je/p
ответилаF	odgovorila/je/F-F	odpowiedziała/F-F	je/o
обиделасьF	,/uvrijedila/F-F	obraziła/F-F	vide
осведомилсяF	raspitivao/se/F-F	zasięgnął/informacji/F-F	vpra
привезутF	će/dopremiti/--	przywiozą/F-F	prip
ответилаF	odgovorila/je/F-F	odpowiedziała/F-F	je/o

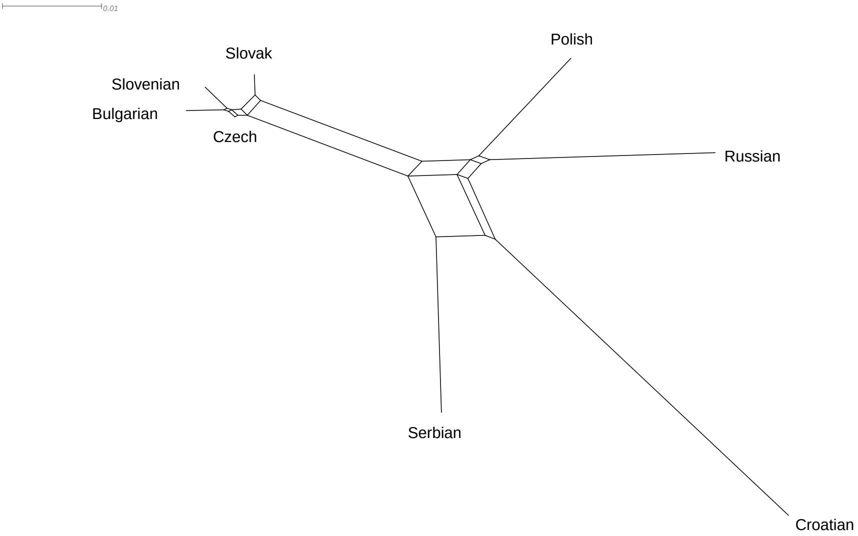
Setup

- past/present active participle / adverbial converbs on the basis of morphological tags, lemmata and word form information
- more inter-text variation expected: run analysis for several books, translated from different languages

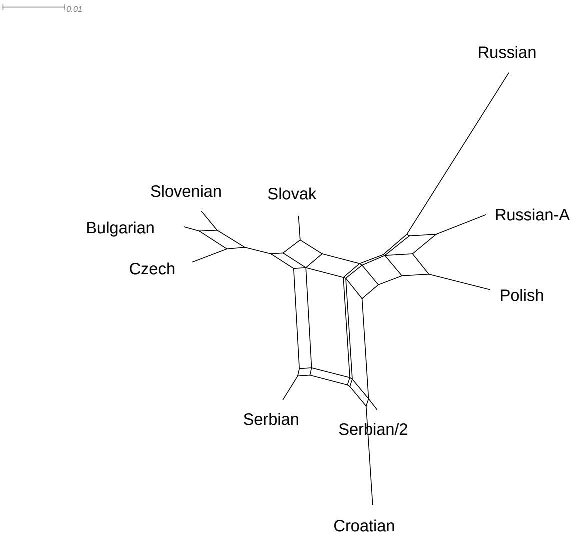
Bulgakov, Master i Margarita, participles vs. finite structures



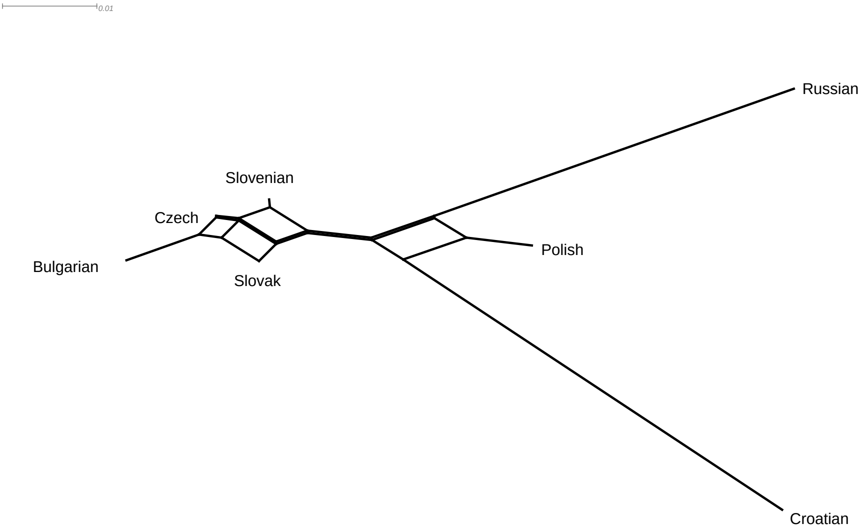
Rowling, Harry Potter, participles vs. finite structures



Lem, Solaris, participles vs. finite structures



Eco, Il nome della rosa, participles vs. finite structures



Conclusions

- very different distribution; Russian, Polish, BCS vs. rest
- still to be worked out

Outline

1. Introduction: Aims and methodology
   - Approaching Slavic variety from a usage point of view
   - ParaSol: A Parallel Corpus of Slavic and other Languages
2. Method
   - A sample variable: Aspect in the imperative
   - Automating procedures
3. Further variables
   - Verbal prefixes
   - Nominal suffixes
   - Participles vs. finite clauses
4. Conclusions

Summary / work plan

Summary and conclusions

- bottom-up approach to *functional* similarity of linguistic variables ranging from grammatical categories via derivational morphology to syntactic preferences
- discovery method: both bird's eye view and access to detail
- some interesting results (aspect distribution, verbal vs. nominal derivation, contact vs. inheritance)

Plan for coming months:

- publish software (?)
- publish selected studies in separate
- put together monography
- ...

Outline

5 Appendix: Further projects

- Language choice in the post-soviet sphere: twitter as a source
- Dialect Corpus: the Language of the Ustja River bassin
- Diachronic corpus interface

Further projects

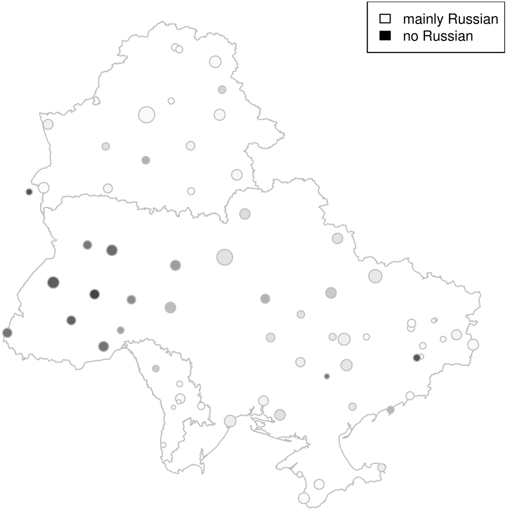
Some more projects I am engaged in and would like to work on while at IPI PAN – approach me if you want to know more!

Outline

5 Appendix: Further projects

- Language choice in the post-soviet sphere: twitter as a source
- Dialect Corpus: the Language of the Ustja River bassin
- Diachronic corpus interface

Language choice in Twitter in Ukraine/neighboring countries



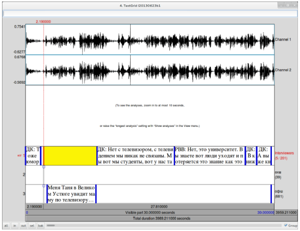
- with Tom Ruette, Berlin; tweet collection since 2011
- simple criteria (alphabet and word list) to detect Russian vs. non-Russian cyrillic text
- map: color for Russianness, size for number of teeters
- pretty good overlap with survey-based results
- project under revitalization

Outline

5 Appendix: Further projects

- Language choice in the post-soviet sphere: twitter as a source
- Dialect Corpus: the Language of the Ustja River bassin
- Diachronic corpus interface

Language of the Ustja River bassin



(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

ДК: Т оже юмор

ДК: Нет с телевизором, с телевидением мы никак не связаны. Мы вот мы студенты, вот у нас та

РВВ: Нет, это университет. Вы знаете вот люди уходят и потеряется это знание как это

ДК: В к ниж

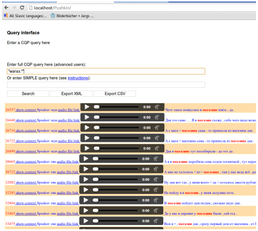
ДК: А вы же кн

Меня Таня в Великом Устюге увидит маму по телевизору…

2.190000 | 27.810000

0 Visible part 30.000000 seconds 30.000000 | 3959.211000

Total duration 3959.211000 seconds



Query interface

Enter a CQP query here

Enter full CQP query here (advanced users):

"магаз.*"

Or enter SIMPLE query here (see instructions):

26557 show context Speaker: мдн audio file link 0:00 Чего такое повкуснее в магазине взять - да .

26648 show context Speaker: мдн audio file link 0:00 Дак это само . . . Я в магазин схожу , себе чего надо возь

26716 show context Speaker: мдн audio file link 0:00 А с мага = магазина сама - то принесла из магазина дак .

26722 show context Speaker: мдн audio file link 0:00 А с мага = магазина сама - то принесла из магазина дак .

26899 show context Speaker: мдн audio file link 0:00 Дак в магазине тут пособирали - да это да .

30665 show context Speaker: еве audio file link 0:00 А я в магазине поробила семь годов техничкой , тут пере

30722 show context Speaker: еве audio file link 0:00 А мне не хотелось < из > магазина , там у нас ведь всё , ра

32082 show context Speaker: еве audio file link 0:00 Ну дак вот где , у меня всего < да > осталось двести рубле

32283 show context Speaker: еве audio file link 0:00 Не пойду я в магазин , у меня дедушко есть .

32864 show context Speaker: еве audio file link 0:00 В магазин пойдет дам подам , сколько надо дак .

33465 show context Speaker: еве audio file link 0:00 Да у нас в деревне у магазина были , сей год .

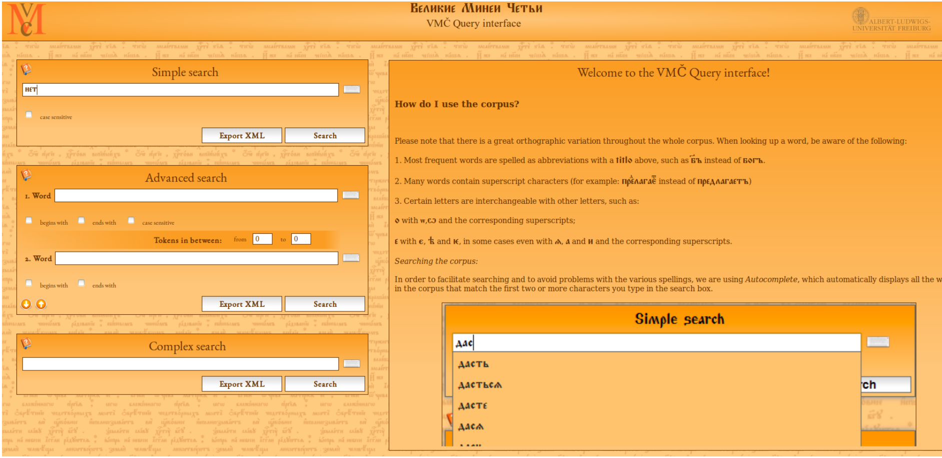
33475 show context Speaker: еве audio file link 0:00 Вон в = , магазин дак , сразу первый дом от магазина , от

- joint project with Nina Dobrushina, Michael Daniel (Higher School of Economics, Moscow)
- data collected in Swiss-Russian field trips to South of Arkhangel'sk region (students and faculty)
- transcribed in **standard Russian**, lemmatized and tagged - already more than 100 000 tokens
- available **with aligned sound segments** via CWB-based web interface; user correction planned to enable **crowdsourcing of the transcription**
- would be **great to cooperate** with other dialect projects!

Outline

5. Appendix: Further projects
   - Language choice in the post-soviet sphere: twitter as a source
   - Dialect Corpus: the Language of the Ustja River bassin
   - Diachronic corpus interface

VMČ - prototype



Simple search: нет

Tokens in between: from 0 to 0

Welcome to the VMČ Query interface!

How do I use the corpus?

Please note that there is a great orthographic variation throughout the whole corpus. When looking up a word, be aware of the following:

1. Most frequent words are spelled as abbreviations with a titlo above, such as б҃ъ instead of богъ.

2. Many words contain superscript characters (for example: пр҃ѣлагаѥ instead of предлагаетъ)

3. Certain letters are interchangeable with other letters, such as:

о with ѡ,ꙍ and the corresponding superscripts;

ɛ with є, ѣ and ѥ, in some cases even with ѧ, ꙗ and и and the corresponding superscripts.

Searching the corpus:

In order to facilitate searching and to avoid problems with the various spellings, we are using *Autocomplete*, which automatically displays all the w
in the corpus that match the first two or more characters you type in the search box.

дас

дасть
дастьсѧ
дастє
дасѧ

- **easy to use** interface to CWB-based corpus of **Velikie minei čet'i**, edited at Slavic department at Freiburg University, Germany

Thanks!

Dziękuję za uwagę!

waldenfels@issl.unibe.ch
ruprecht.waldenfels@gmail.com