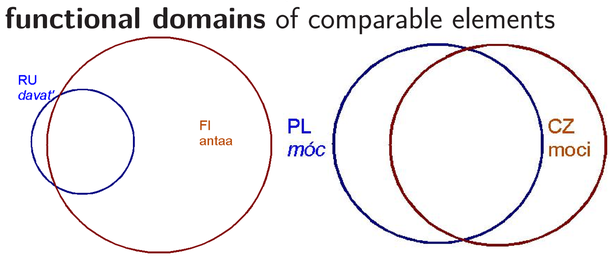
Reconstructing functional change based on a parallel corpus: The expansion of DO in Northern Slavic

Ruprecht von Waldenfels, Institute of Polish, Polish Academy of Sciences, Cracow
Slavic Corpus Linguistics: The Historical Dimension Tromsø, Norway, April 21–22, 2015

Comparing functional domains using a parallel corpus

Distributional semantics across languages

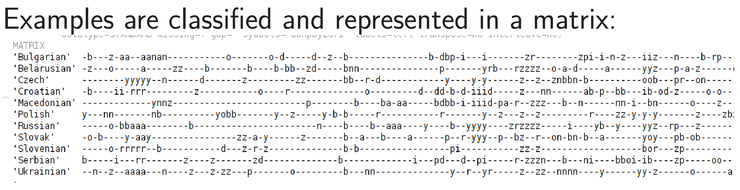
Distribution in parallel (translated) texts as a proxy for **functional domains** of comparable elements



ParaSol corpus: 32 languages, >400 pairs, 25 mio token; 8 novels in all major Slavic standard languages

Aggregating over many examples

Examples are classified and represented in a matrix:



The more classifications coincide across languages, the more similar these languages are held to be in respect to feature in question (Hamming distance).

The resulting distances are visualized in NeighborNets (see below).

Operationalization

The classification is defined in parameter files based on lemmatization, POS-tagging and word alignment:

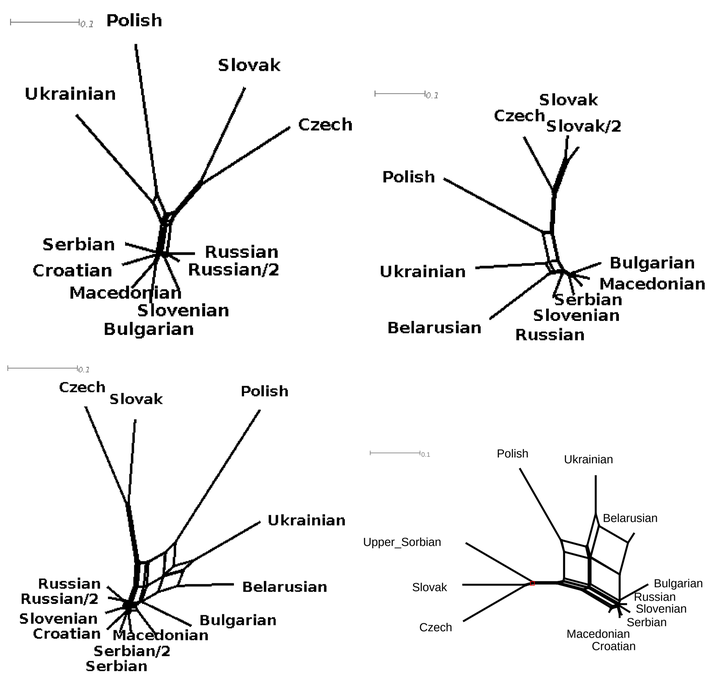
```
<type id="N" name="NA">
        <criteria><lng>ru</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>uk</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^S.*</regexp>   </criteria>
        <criteria><lng>by</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^S.*</regexp>   </criteria>
        <criteria><lng>mk</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^[Nn].*</regexp>   </criteria>
        <criteria><lng>bg</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>pl</lng><regexp level="preclem">^на$</regexp><regexp level="tag">^subst.*</regexp></criteria>
        <criteria><lng>cz</lng><regexp level="preclem">^na(-|d)?$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>sk</lng><regexp level="preclem">^na$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>sl</lng><regexp level="preclem">^na$</regexp><regexp level="tag">^S.*</regexp>   </criteria>
        <criteria><lng>hr</lng><regexp level="preclem">^na$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>sr</lng><regexp level="preclem">^na$</regexp><regexp level="tag">^N.*</regexp>   </criteria>
        <criteria><lng>de</lng><regexp level="preclem">^aufs?$</regexp><regexp level="tag">^NN$</regexp>   </criteria>
        <criteria><lng>lt</lng><regexp level="preclem">^ant$</regexp><regexp level="tag">^dktv:</regexp>   </criteria>
        <criteria><lng>us</lng><regexp level="precword">^[nN]a$</regexp>   </criteria>

 </type>
```

http://www.parasolcorpus.org

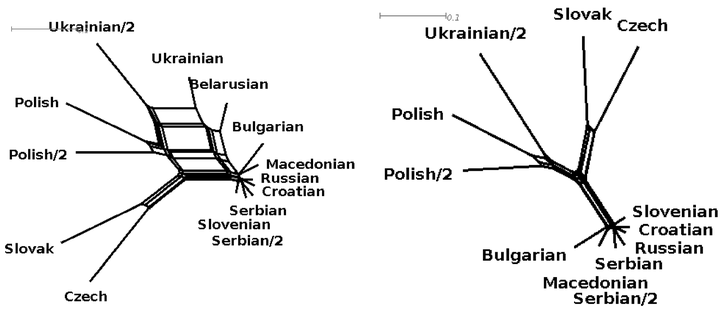
Distribution of prepositions across languages: quantitative and qualitative perspectives

Visualization of similarity in use of DO

	BULGAKOV			ROWLING			LEM		
	DO	K	Kon	DO	K	Kon	DO	K	Kon
ru	83	417	0	60	233	0	43	127	0
rua	--	--	--	65	208	0	45	112	0
by	403	8	0	--	--	--	179	0	0
uk	684	2	0	406	0	0	280	0	0
uka	1038	2	0	--	--	--	--	--	--
pl	891	65	0	472	21	0	308	18	0
pla	714	114	0	--	--	--	--	--	--
cz	702	372	0	342	226	0	226	140	0
sk	719	320	0	344	156	0	206	87	0
ska	--	--	--	--	--	--	--	--	--
sl	85	273	0	88	108	0	48	50	0
hr	108	70	210	113	5	46	59	1	9
sr	137	77	143	97	47	36	81	31	39
sra	140	3	200	--	--	--	69	16	11
mk	0	0	329	0	0	188	0	0	108
bg	198	0	546	120	0	282	99	0	148



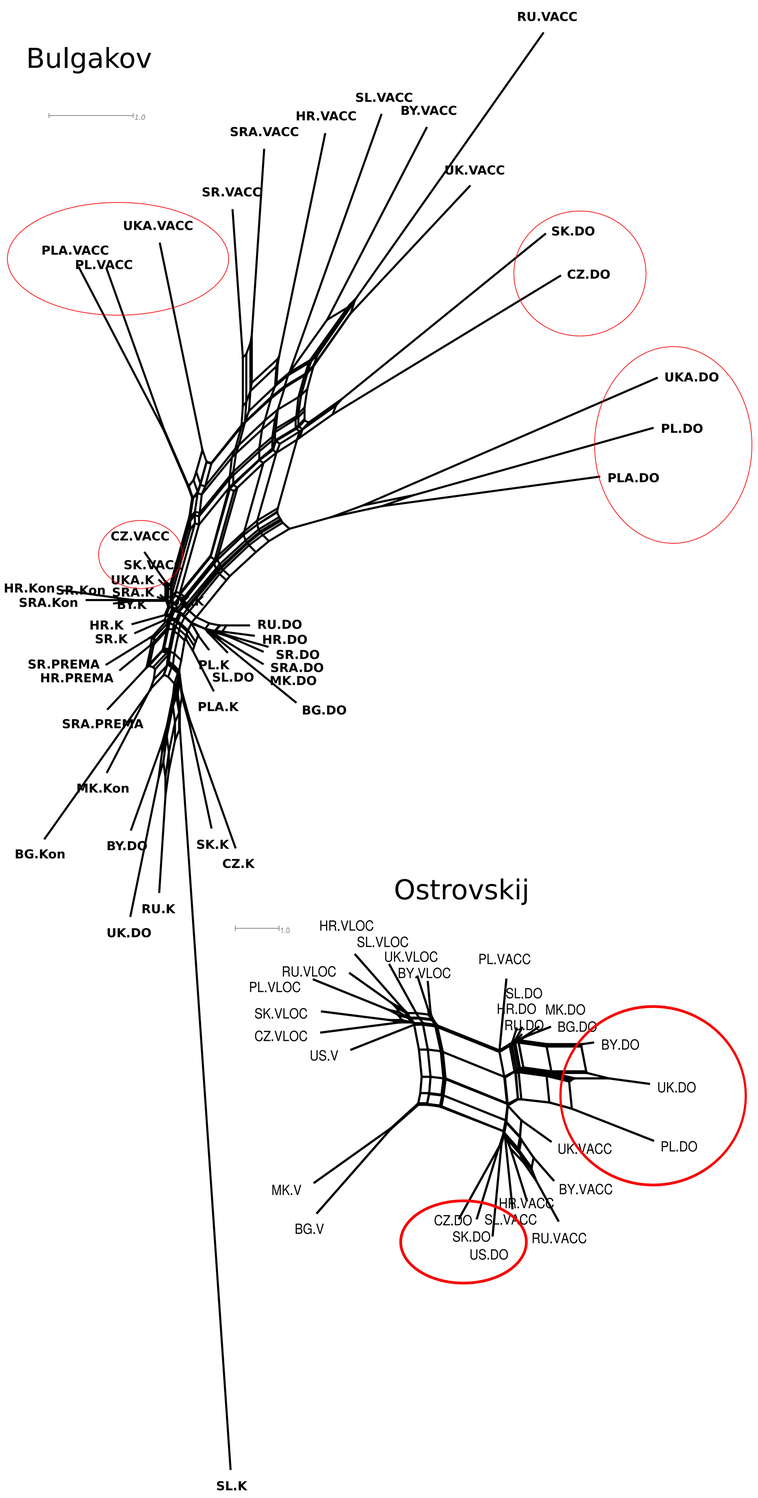
From upper left to lower right: Rowling, *Harry Potter and the Sorcerer's Stone*; Andrić, *Na Drini ćuprija*; Lem,*Solaris*; Ostrovskij, *Kak zakaljalos' stal'*.



Bulgakov, *Master i Margarita* with conflicting affiliations of two Ukrainian versions; if Belarusian and Ukrainian-1 are removed (right graph), Ukrainian/2 forms forms a close group with Polish.

Inspecting language-specific formants

- Slovak/Czech/Sorbian DO in one cluster with V-ACC
- Polish/Ukrainian DO intermediate
- Belarusian/Ukrainian DO in one cluster with K
- Two types of Ukrainian translations



DO in Czech/Slovak/Sorbian, but not in Polish: typically V-ACC used in Polish and rest of Slavic:

(1) *Milan Kundera, Nesnesitelná lehkost bytí*

- cz Pohnut tou představou, vtiskl v té chvíli tvář do polštáře vedle její hlavy a dlouho tak zůstal.
- ru Rastrogannyj ėtim voobražaemym obrazom, on zarylsja licom v podušku [..]
- uk Sxvil'ovanij cijeju dumkoju, Tomaš utknuvsja obliččjam u podušku [..]
- pl Poruszony tą wizją wcisnął twarz w poduszkę obok jej głowy [..]
- sl Pretresen nad to podobo je zakopal obraz v blazino poleg njene glave in dolgo ostal v tem položaju.
- hr U tom času utisnuo je lice u jastuk pored njene glave [..]
- mk [..] go pritisna liceto vo pernicata kraj nejzinata glava [..]
- bg Raznežen ot tazi misăl, zarovi lice văv văzglavnicata do liceto na Tereza

DO in one, not the other Ukrainian translation likewise the equivalence of DO and V(ACC)

(2) *Bulgakov, Master i Margarita*

- ru Ivan [. . .] ne uderžalsja i, vidja, kak voda xleščet v vannu širokoj struej iz sijajuščego krana, skazal s ironiej:
- by . . . šyrokim strumenem prė ŭ vannu vada . . .
- uk . . . jak voda rine u vannu širokim strumenem . . .
- uka . . . jak voda plešče do vanni širokim strumenem . . .
- pl . . . obserwując wodę [. . .] lejącą się [. . .] do wanny . . .
- pla . . . woda szeroko chlusta ze lśniącego kranu do wanny . . .
- cz . . . který se řinul z nablýskaného kohoutku do vany . . .
- sk . . . širokým prúdom tečie do vane voda . . .
- sl . . . kako se voda v širokem curku zliva [. . .] v kad . . .
- hr . . . iz blistave pipe voda u širokoj struji teče u kadu . . .
- sr . . . voda juri u kadu širokim mlazom iz blistave slavine . . .
- sra . . . kako se voda iz blistave slavine [. . .] izliva u kadu . . .
- bg . . . kak vodata pljušti văv vanata na široka struja . . .

Belarusian and Ukrainian, to a lesser degree Polish, use DO for Russian and Czech K

(3)
- ru Prygajuščej rukoj podnes Stepa stopku k ustam, [. . .].
- by Dryžačaj rukoju padnës Scëpa garėlku da vusnaŭ, [. . .].
- uk Nepevnoju rukoju pidnis St'opa do vust čarku, [. . .].
- uka Tremtjačoju rukoju pidnis St'opa stopku do vust, [. . .].
- pl Drżącą ręką Stiopa podniósł kieliszek do ust, [. . .].
- pla Stiopa podniósł rozdygotaną ręką szklaneczkę do ust, [. . .].
- cz Lotrov rozechvělou rukou pozvedl sklenici k ústům [. . .].
- sk Stopa roztrasenou rukou zdvihol k ústam pohárik [. . .].
- sl « S tresočo se roko je Stjopa ponesel kozarec k ustom, [. . .].
- hr Stjopa je drhtavom rukom prinio čašicu **ustima**, [. . .].
- sr Drhtavom rukom Stjopa prinese čašicu **ustima**, [. . .].
- sra So rastreperena raka Stjopa ja dobliži čašata do usta
- bg St'opa podnese s trepereša răka čašata kăm ustata si

Results: two directions of functional extension of DO in North Slavic

Schematic distribution of prepositional functions

RU	BY	PL	CZ	US
K	DO	K	K	K
		DO		
DO			DO	DO
V	V			
		V		
			V	

- **DO for V(ACC): Sorbian>Czech/Slovak>Polish>Ukrainian**
- **DO for K: Belarusian/Ukrainian > Polish**

Results

- two opposing changes proceeding from West and East
- mentioned in Kopečný (1973) in passing, but much more detail here
- origin of this change in Sorbian remains to be explained; no plausible contact influence of German
- variation in Ukrainian translations probably reflects contact influence and finds parallels in Ukrainian standard variation

Acknowledgement

Funding by the Swiss National Science foundation (project *Convergence and divergence of Slavic from a usage based, parallel corpus driven perspective*) is gratefully acknowledged.

References

Huson, D. H. and D. Bryant (2006): Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.*, 23(2):254-267

Kopečný, František (1973). *Etymologický slovník slovanských jazyku. Slova gramatická a zájmena. Svazek 1: Předložky, koncové partikule.* Praha: Academia

Vaillant, A. (1950-1979). *Grammaire comparée des langues slaves.* Paris: Institute d'étude slaves

v. Waldenfels, Ruprecht (2014): Explorations into variation across Slavic: taking a bottom-up approach. In: Benedikt Szmrecsanyi, Bernhard Wälchli (ed.): *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages.* De Gruyter Mouton