

The 47<sup>th</sup> meeting of the commission on the grammatical structure  
of Slavic languages of the International committee of Slavists  
Belgrade, September 12-14, 2024

# Bivalent verb classes across Slavic: areal and genealogical patterns

Sergey Say  
[serjzhka@yahoo.com](mailto:serjzhka@yahoo.com)  
University of Potsdam



# Intro

## Serbian

*Petar*      *se*      *stidi*      *svoj-e*      *visin-e*  
PN.NOM.SG      REFL      shame.PRS.3SG      one's-GEN.SG.F      height-GEN.SG

‘Petar is embarrassed about his height.’

## Czech

*Petr*      *se*      *stydí*      *za*      *svoj-i*      *mal-ou*      *postav-u*  
PN.NOM.SG      REFL      shame.PRS.3SG      for      one's-F.ACC.SG      small-F.ACC.SG      stature(F)-ACC.SG

‘Petr is embarrassed about his small height.’

# Intro

	Verb	Pattern
Serbian	<i>stiditi se</i>	NOM_GEN
Czech	<i>stydět se</i>	NOM_zACC

These two valency patterns in Serbian and Czech...

- are parts of their respective systems
  - are both partly motivated by the verbs' meanings
  - are used with cognate verbs
  - are not cognate to each other
  - are based on different cognitive schemas
- => Divergence

# Intro: Goals

- Quantitatively assess similarities and differences between Slavic languages in the domain of valency encoding
- Compare lexical vs. syntactic dimension
- Interpret results in the genealogical and areal dimensions

# Roadmap

- Valency classes: basic ideas
- Dataset
  - BivaTyp
  - additional annotation for Slavic
- Distance metrics
- Results
- Summary

# Valency classes: basic ideas

- The valency of a verb = “the list of its arguments with their coding properties” (Malchukov et al. 2015: 30)
- Coding properties (devices)
  - flagging: cases & adpositions => **relevant for Slavic!**
  - indexing: agreement, cross-referencing
  - word order (rarely)

# Valency classes: basic ideas

- Three Serbian examples

*Petar*      *liči*      *na*      *Marij-u*  
PN.NOM.SG      resemble.PRS.3SG      on      PN-ACC.SG

‘Petar **resembles** Maria.’

*moj-e*      *ruk-e*      *miriš-u*      *na*      *benzin*  
my-NOM.PL.F      hand-NOM.PL      smell-PRS.3PL      on      gasoline.ACC.SG

‘My hands **smell** of gasoline.’

*Petar*      *se*      *ljuti*      *na*      *Marij-u*  
PN.NOM.SG      REFL      anger.PRS.3SG      on      PN-ACC.SG

‘Petar is **angry** with Maria.’

=> These three sentences represent **the same** valency pattern in Serbian: **NOM\_naACC**

# Valency classes: basic ideas

- Three sources for valency classes (aka “case assignment”)
  - *Syntax* (“*structural case*”)
  - *Semantics* (“*thematic case*”)
  - *Lexicon* (“*lexical case*”)
- Empirical questions
  - how semantically motivated are valency classes?
  - how stable are they diachronically?
  - what happens when a verb is lexically renewed?



# Valency classes: basic ideas

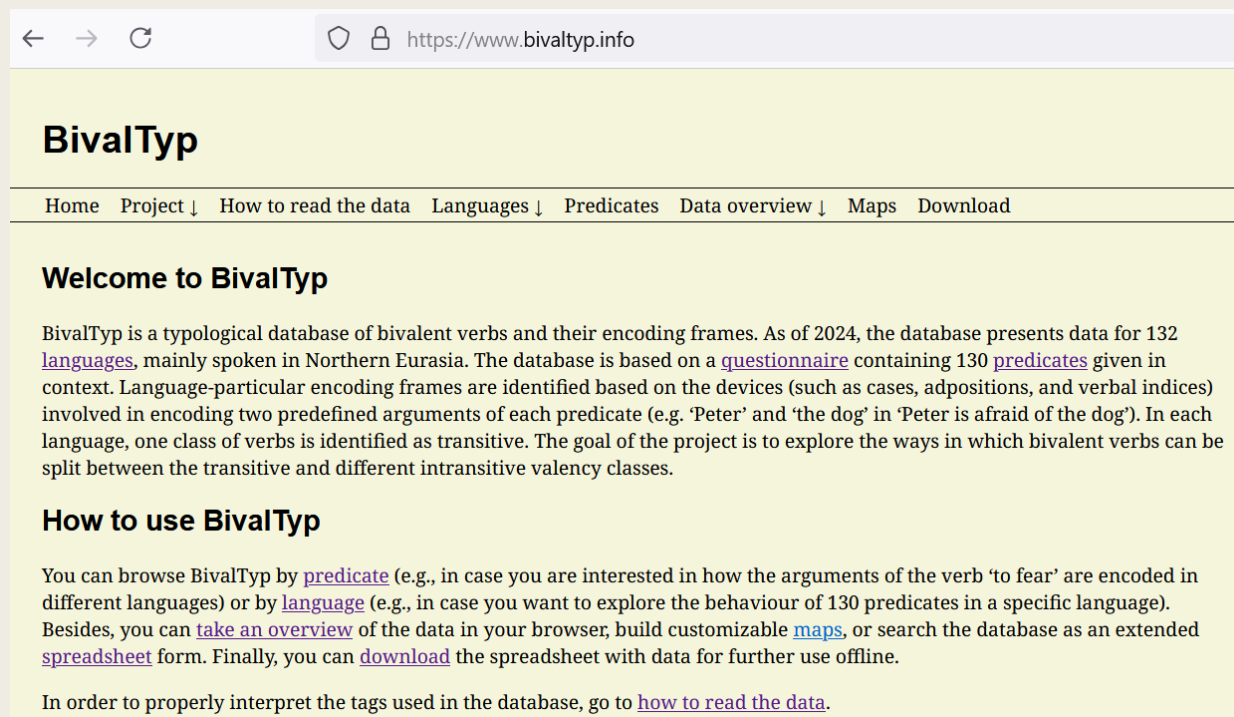
- Bivalent verbs (= verbs with two arguments) are especially prone to display deviant (=non-canonical) valency behavior (Bickel et al. 2014), unlike monovalent verbs
- Bivalent verbs often form relatively large classes, unlike non-canonical trivalent verbs

# Dataset: Bivaltyp

Sergey Say (ed.). 2020–... BivalTyp: Typological database of bivalent verbs and their encoding frames.

St. Petersburg: Institute for Linguistic Studies, RAS.

(Available online at <https://www.bivaltyp.info>)



The screenshot shows the homepage of the BivalTyp website. At the top, there is a navigation bar with links: Home, Project ↓, How to read the data, Languages ↓, Predicates, Data overview ↓, Maps, and Download. Below the navigation bar, the title "BivalTyp" is displayed. The main content area starts with a "Welcome to BivalTyp" section, followed by a paragraph describing the database: "BivalTyp is a typological database of bivalent verbs and their encoding frames. As of 2024, the database presents data for 132 languages, mainly spoken in Northern Eurasia. The database is based on a questionnaire containing 130 predicates given in context. Language-particular encoding frames are identified based on the devices (such as cases, adpositions, and verbal indices) involved in encoding two predefined arguments of each predicate (e.g. 'Peter' and 'the dog' in 'Peter is afraid of the dog'). In each language, one class of verbs is identified as transitive. The goal of the project is to explore the ways in which bivalent verbs can be split between the transitive and different intransitive valency classes." Below this is a "How to use BivalTyp" section, which states: "You can browse BivalTyp by predicate (e.g., in case you are interested in how the arguments of the verb 'to fear' are encoded in different languages) or by language (e.g., in case you want to explore the behaviour of 130 predicates in a specific language). Besides, you can take an overview of the data in your browser, build customizable maps, or search the database as an extended spreadsheet form. Finally, you can download the spreadsheet with data for further use offline." At the bottom of the page, there is a note: "In order to properly interpret the tags used in the database, go to how to read the data."

← → ↺ <https://www.bivaltyp.info>

## BivalTyp

Home Project ↓ How to read the data Languages ↓ Predicates Data overview ↓ Maps Download

### Welcome to BivalTyp

BivalTyp is a typological database of bivalent verbs and their encoding frames. As of 2024, the database presents data for 132 [languages](#), mainly spoken in Northern Eurasia. The database is based on a [questionnaire](#) containing 130 [predicates](#) given in context. Language-particular encoding frames are identified based on the devices (such as cases, adpositions, and verbal indices) involved in encoding two predefined arguments of each predicate (e.g. 'Peter' and 'the dog' in 'Peter is afraid of the dog'). In each language, one class of verbs is identified as transitive. The goal of the project is to explore the ways in which bivalent verbs can be split between the transitive and different intransitive valency classes.

### How to use BivalTyp

You can browse BivalTyp by [predicate](#) (e.g., in case you are interested in how the arguments of the verb 'to fear' are encoded in different languages) or by [language](#) (e.g., in case you want to explore the behaviour of 130 predicates in a specific language). Besides, you can [take an overview](#) of the data in your browser, build customizable [maps](#), or search the database as an extended [spreadsheet](#) form. Finally, you can [download](#) the spreadsheet with data for further use offline.

In order to properly interpret the tags used in the database, go to [how to read the data](#).

# Dataset: BivaTyp

- First-hand data provided by language experts
  - St. Petersburg-style typology
- Questionnaire with 130 verbs given in context
  - Wordlist-based approach: Nedjalkov 1969, Bossong 1998, Nichols et al. 2004, Nichols 2008, Malchukov & Comrie (eds.) 2015, etc.

# Dataset: Bivaltyp

## #21 (Peter was crossing the river in a boat)

'Peter **reached** the bank'

X \_\_\_\_\_ Y

## #22 (The wall was covered with fresh paint)

‘Peter **touched** the wall’ (and got dirty)

X \_\_\_\_\_ Y

=> Two pre-defined arguments (X, Y) for each predicate

# Dataset: Bivaltyp

- Valency classes are language specific
  - identical labels in different languages can represent very different classes
  - similar classes in different languages can have different labels

Abaza (< Northwest Caucasian)

*l-an*                      *zaréma*                      *də-l-c-qraʕa-d*

3SG.F.IO-mother    PN

3SG.H.ABS-3SG.F.IO-COM-help(AOR)-DCL

‘Mother helped Zarema’

=> ABS\_COM

Aghul (Nakh-Daghestanian)

*aslan*                      *meHemed.i-qaj*                      *uqː.a-a*

PN[ABS]

PN-COM

fight.IPF-PRS

‘Aslan is fighting with Muhammad.’

=> ABS\_COM

# Dataset: Bivaltyp

- In BivalTyp, a verb is considered transitive if its two core arguments are coded like the ‘killer’ and the ‘victim’ micro-roles of the ‘kill’ verb (cf. Haspelmath 2015)

- Russian

#105:      *Петя убил Машу*                      => **TR** (by definition)  
              ‘Petja killed Maša’

#28        *Петя ждет Машу*                      => **TR**  
              ‘Petja is waiting for Maša’

#21        *Петя достиг берега*                      => **NOM\_GEN**  
              ‘Petja reached the bank’

# Dataset: Bivaltyp

- The sample: currently 132 languages, mainly spoken in Northern Eurasia



# Dataset: Bivaltyp

## 11 Standard Slavic languages in the dataset

- Russian (Sergey Say)
- Belarusian (Olga Gorickaja)
- Ukrainian (Natalia Zaika)
- Polish (George Moroz)
- Czech (Anastasia Makarova)
- Slovak (Martin Gális, Kirill Kozhanov)
- Slovenian (Andreja Žele, Mladen Uhlik)
- Croatian (Mislav Benić)
- Serbian (Anastasia Escher)
- Macedonian (Vladimir Fedorov, Maria Khazhomia)
- Bulgarian (Krasimira Petrova)



# Additional annotation for Slavic

Verb cognacy sets: if and only if roots (!) are cognate

- RU    *Петя **наказал** своего сына*
- BR    *Алесь **пакараў** свайго сына*
- UK    *Петро **покарав** свого сина*
- PL    *Andrzej **ukarał** swego syna*
- CZ    *Petr **potrestal** syna*
- SK    *Peter **potrestal** svojho syna*
- SL    *Peter je **kaznoval** sina*
- HR    *Pero je **kaznio** sina*
- SR    *Петар је **казнио** свог сина*
- MK    *Петар го **казни** својот син*
- BG    *Петър **наказа** сина си*

# Additional annotation for Slavic

Valency encoding devices cognacy sets

- for case categories (NOM, ACC, GEN, DAT, INS, LOC)
  - etymological relatedness across languages with cases
  - for languages without nominal case (Bulgarian, Macedonian): pronouns are taken into account
- for adpositions
  - etymological relatedness across Slavic
  - in case of mergers, the rule of thumb is to not annotate for differences whenever this is plausible
  - etc. (some complicated scenarios)

# Additional annotation for Slavic

Valency encoding devices cognacy sets

RU	<i>Петя дотронулся до стены</i>	NOM_doGEN
BR	<i>Алесь дакрануўся да сцяны</i>	NOM_doGEN
UK	<i>Петро доторкнувся до стіни</i>	NOM_doGEN
PL	<i>Marek dotknął ściany</i>	NOM_GEN
CZ	<i>Petr se dotkl stěny</i>	NOM_GEN
SK	<i>Peter sa dotkol steny</i>	NOM_GEN
SL	<i>Peter se je dotaknil stene</i>	NOM_GEN
HR	<i>Pero je dodirnuo zid</i>	TR
SR	<i>Петар је додирнуо зид</i>	TR
MK	<i>Петар се допре до сидот</i>	NOM_doGEN
BG	<i>Петър се допря до стената</i>	NOM_doGEN

# Additional annotation for Slavic

NB! Verbs' and valency encoding devices' cognacy relations are logically and empirically independent of each other!

RU	<i>Петя <b>забыл</b> о другой дороге</i>	<b>NOM_oLOC</b>
PL	<i>Basia <b>zapomniała</b> o tej drodze</i>	<b>NOM_oLOC</b>
CZ	<i>Petr <b>zapomněl</b> na tu druhou cestu</i>	<b>NOM_naACC</b>
SK	<i>Peter <b>zabudol</b> na druhú cestu</i>	<b>NOM_naACC</b>
SL	<i>Peter je <b>pozabil</b> na pot</i>	<b>NOM_naACC</b>

# Distance metrics

- All distance metrics are based on pairwise comparisons between two languages (L1, L2)
- All metrics are normalized, where 0 corresponds to total identity, and 1, to total dissimilarity

	Type of information
DistEtymVerb	How often are the verbs in L1 and L2 not cognate to each other?
DistEtymPat	How often are the valency encoding devices in L1 and L2 not cognate to each other?
DistValLoc	How often do the entries in L1 and L2 diverge in terms of locus of non-transitivity?
DistValPat	How dissimilar are lexical extents of the valency classes in L1 and L2 (set-partitions)?

# Distance metrics: overview

	Scope	Type
DistEtymVerb	Slavic	Normalized Hamming distance
DistEtymPat	Slavic	Normalized Hamming distance
DistValLoc	Universal	Normalized Hamming distance
DistValPat	Universal	based on Mutual Information

# DistEtymVerb

No	Russian	Polish	Non-cognate verb?
42	ЛИШИТЬСЯ	stracić	1
43	ЛОВИТЬ	łapać	1
44	СЛОМАТЬ	złamać	0
45	ЛЬСТИТЬ	pochlebiać	1
46	ЛЮБИТЬ	kochać	1
47	МАХАТЬ	machać	0
48	МЕЧТАТЬ	marzyć	1
49	ВЫМЫТЬ	umyć	0
50	НАДЕТЬ	włożyć	1
51	НАЗЫВАТЬСЯ	nazywać się	0

- With this toy subsample of 10 entries per language, the normalized Hamming distance would equal 0.6 (= 6/10)
- In the dataset, DistEtymVerb (RU, PL) = 0.57 = 74/130

# DistEtymPat

	Ukrainian		Polish		non-cognate?
	Pattern	EtymPat	Pattern	EtymPat	
55	TR	TR	TR	TR	0
56	DAT_GEN	DAT_GEN	DAT_GEN	DAT_GEN	0
57	TR	TR	NOM_GEN	NOM_GEN	1
58	DAT_NOM	DAT_NOM	DAT_NOM	DAT_NOM	0
59	DAT_NOM	DAT_NOM	NOM_GEN	NOM_GEN	1
60	TR	TR	TR	TR	0
61	uGEN_NOM	uGEN_NOM	DAT_NOM	DAT_NOM	1
62	NOM_DAT	NOM_DAT	NOM_DAT	NOM_DAT	0
63	TR	TR	TR	TR	0
64	NOM_vidGEN	NOM_otGEN	NOM_odGEN	NOM_otGEN	0

- With this toy subsample of 10 entries per language, the normalized Hamming distance would equal 0.3 (= 3/10)
- In the dataset, DistEtymPat (UK, PL) = 0.205 = 26/127



# DistValLoc

- This metric is based on the notion of “locus of (non-)transitivity”

Russian

*Пете снится Маша*                      => X-locus

Serbian

*Петар сања Марију*                      => TR

Slovak

*Peter sníva o Márii*                      => Y-locus

# DistValLoc

- This metric is based on the notion of “**locus of (non-)transitivity**”
- Four-way classification of valency patterns based on which of the two arguments – X, Y, both, neither – are encoded as non-core arguments
- NB: the classification represents a comparative concept and is applicable universally!
- Pairwise distances between languages are again calculated as normalized Hamming distances

# DistValPat

- Cross-linguistic identification of minor minor valency classes (cf. “ablative verbs”?, “instrumental verbs”?) is not feasible
- Measuring (dis)similarity in valency class systems is the biggest challenge
- I propose **DistValPat**, a metric based on entropy and MI (mutual information)
- Entropy  $\approx$  the amount of information (conveyed by the valency class assignment)

	Armenian	Azerbaijani	Joint Distribution
take	TR	TR	TR_TR
see	TR	TR	TR_TR
influence	NOMvra	NOMDAT	NOMvra_NOMDAT
encounter	TR	NOMCOM	TR_NOMCOM
enter	NOMNOM	NOMCOM	NOMNOM_NOMCOM
win	TR	NOMDAT	TR_NOMDAT
go_out	NOMABL	NOMABL	NOMABL_NOMABL
drive	TR	TR	TR_TR
bend	TR	TR	TR_TR
tell	NOMDAT	TR	NOMDAT_TR
hold	TR	TR	TR_TR
catch_up	NOMDAT	NOMDAT	NOMDAT_NOMDAT
milk	TR	TR	TR_TR
reach	NOMDAT	NOMDAT	NOMDAT_NOMDAT
touch	NOMDAT	NOMDAT	NOMDAT_NOMDAT
fight	NOMhet	NOMCOM	NOMhet_NOMCOM
be_friends	NOMhet	NOMCOM	NOMhet_NOMCOM
think	NOMmasin	NOMABL	NOMmasin_NOMABL
...			
H (Entropy)	1.658	1.462	2.196

# DistValPat

- MI (Mutual Information) =  $H(X) + H(Y) - H(X, Y)$
- Higher MI values reflect higher similarity between valency class systems in the two languages
- MI was calculated using R package `infotheo` (Meyer 2014)

- Converting MI into a distance metric

$$\text{DistValPat}(L1, L2) = 1 - \frac{\frac{MI(L1, L2)}{H(L1)} + \frac{MI(L1, L2)}{H(L2)}}{2}$$

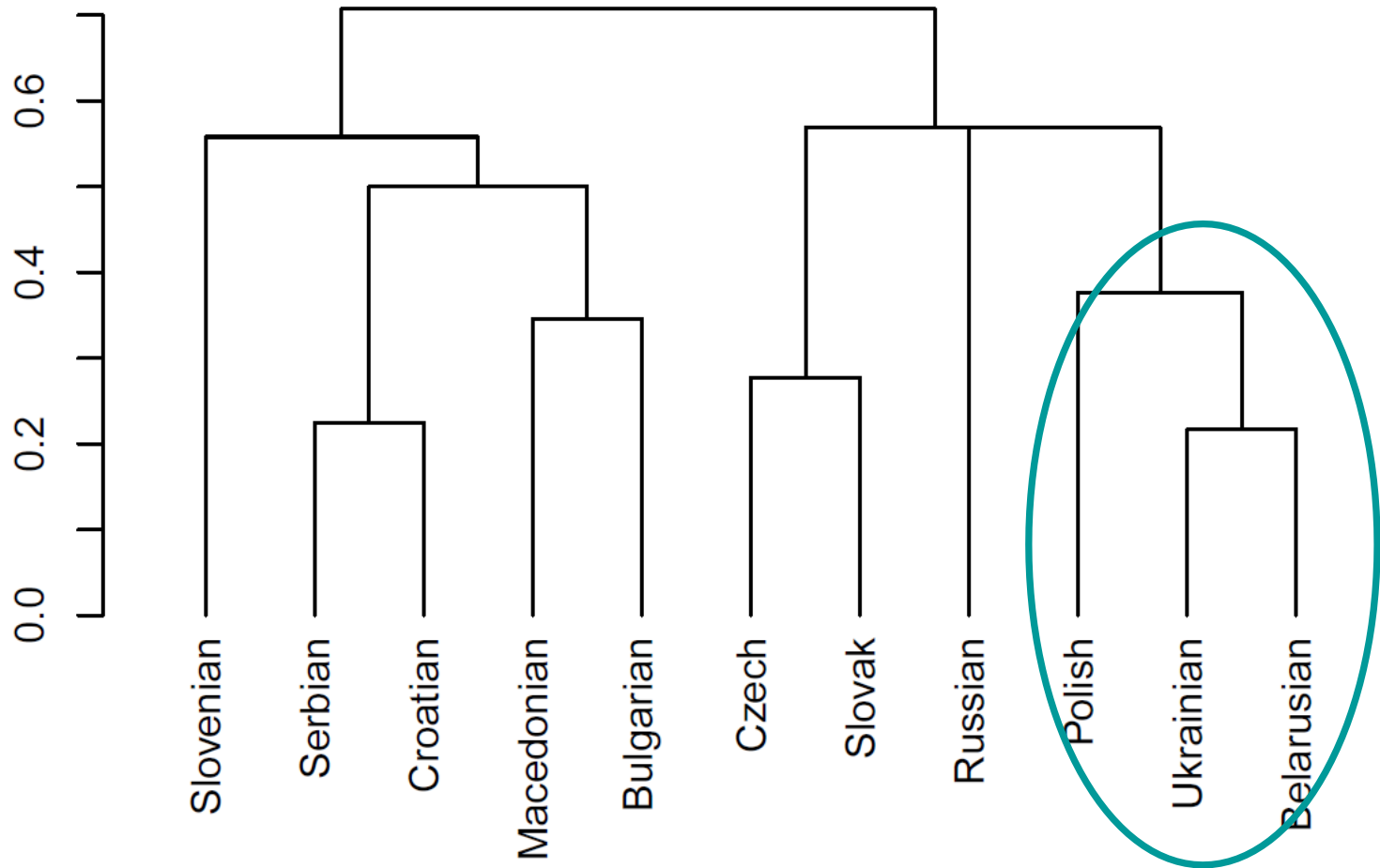
- DistValPat is high if the joint entropy is high relative to individual entropies
- DistValPat is higher if valency class systems are divergent

# Distance metrics: summary

- For each of the 4 distance metrics, I created distance matrices
  - DistEtymVerb: 11 by 11
  - DistEtymPat: 11 by 11
  - DistValLoc: 132 by 132
  - DistValPat: 132 by 132
- Standard methods for dimensionality reduction and visualization
  - Hierarchical clusterization (HClust)
  - Multi-dimensionality Scaling (MDS)
  - NeighborNet (implemented in SplitsTree software)

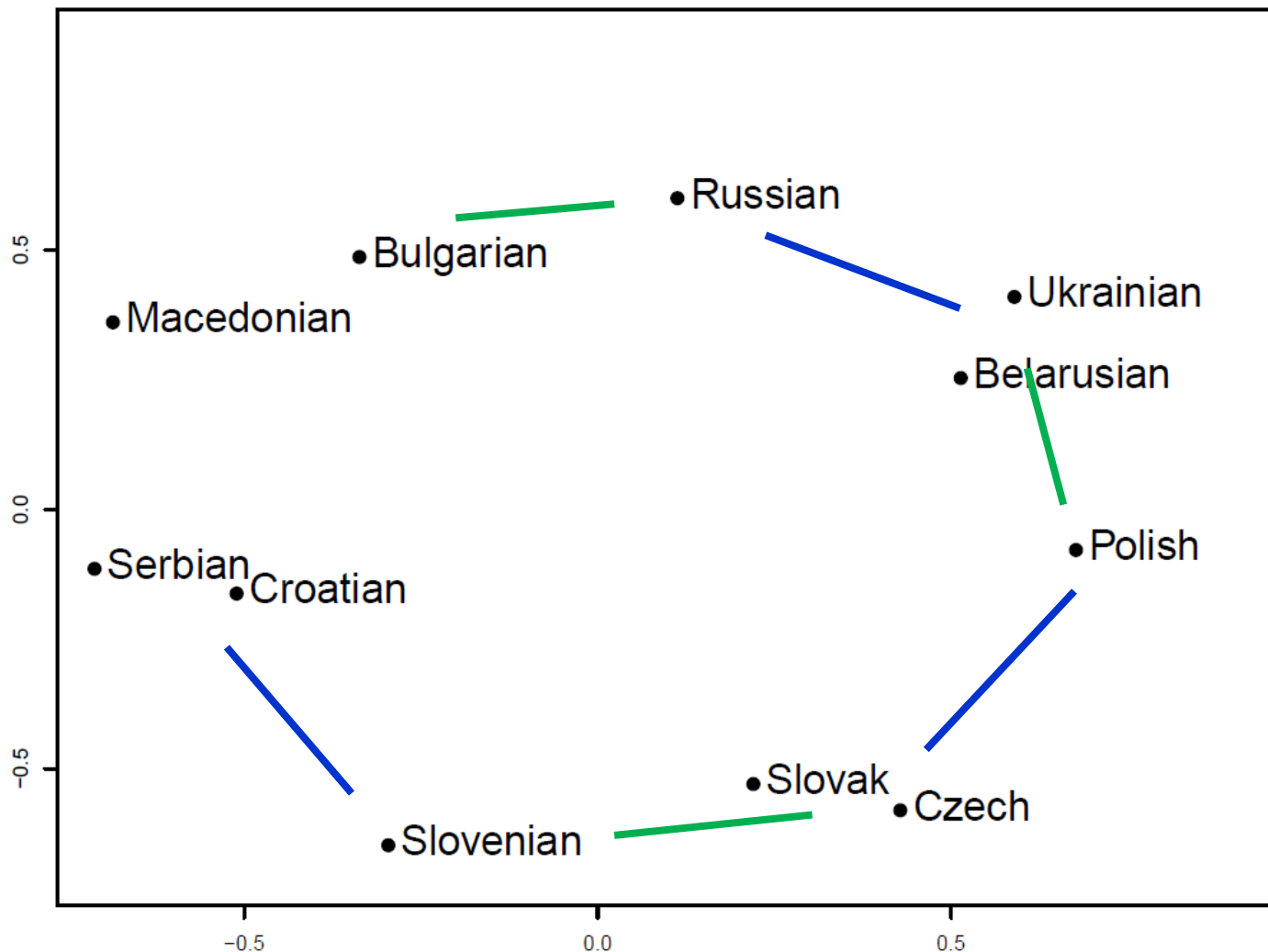
# Results: DistEtymVerb

Hierarchical clusters based on DistEtymVerb



# Results: DistEtymVerb

MDS-visualization based on DistEtymVerb

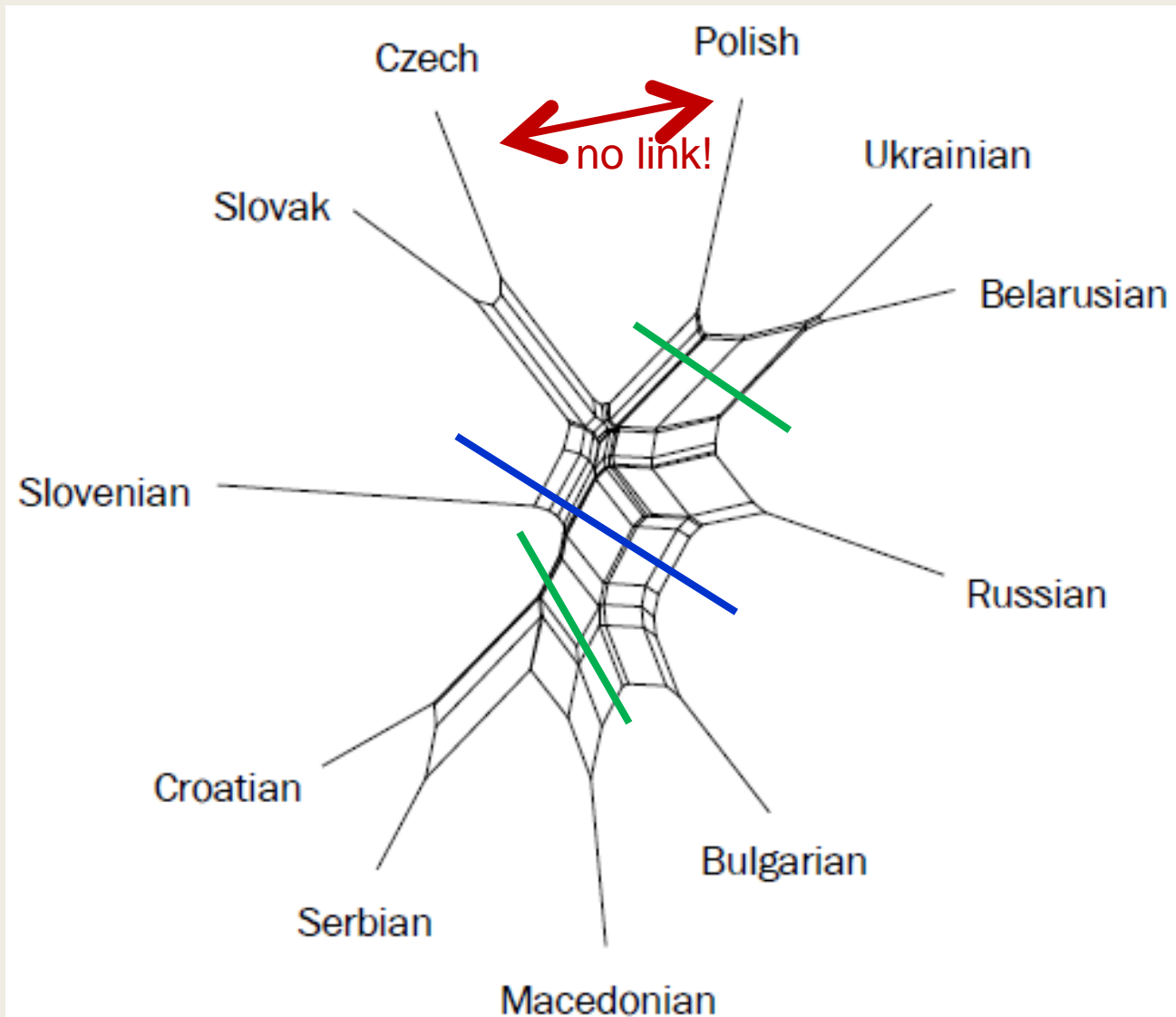


Genealogy

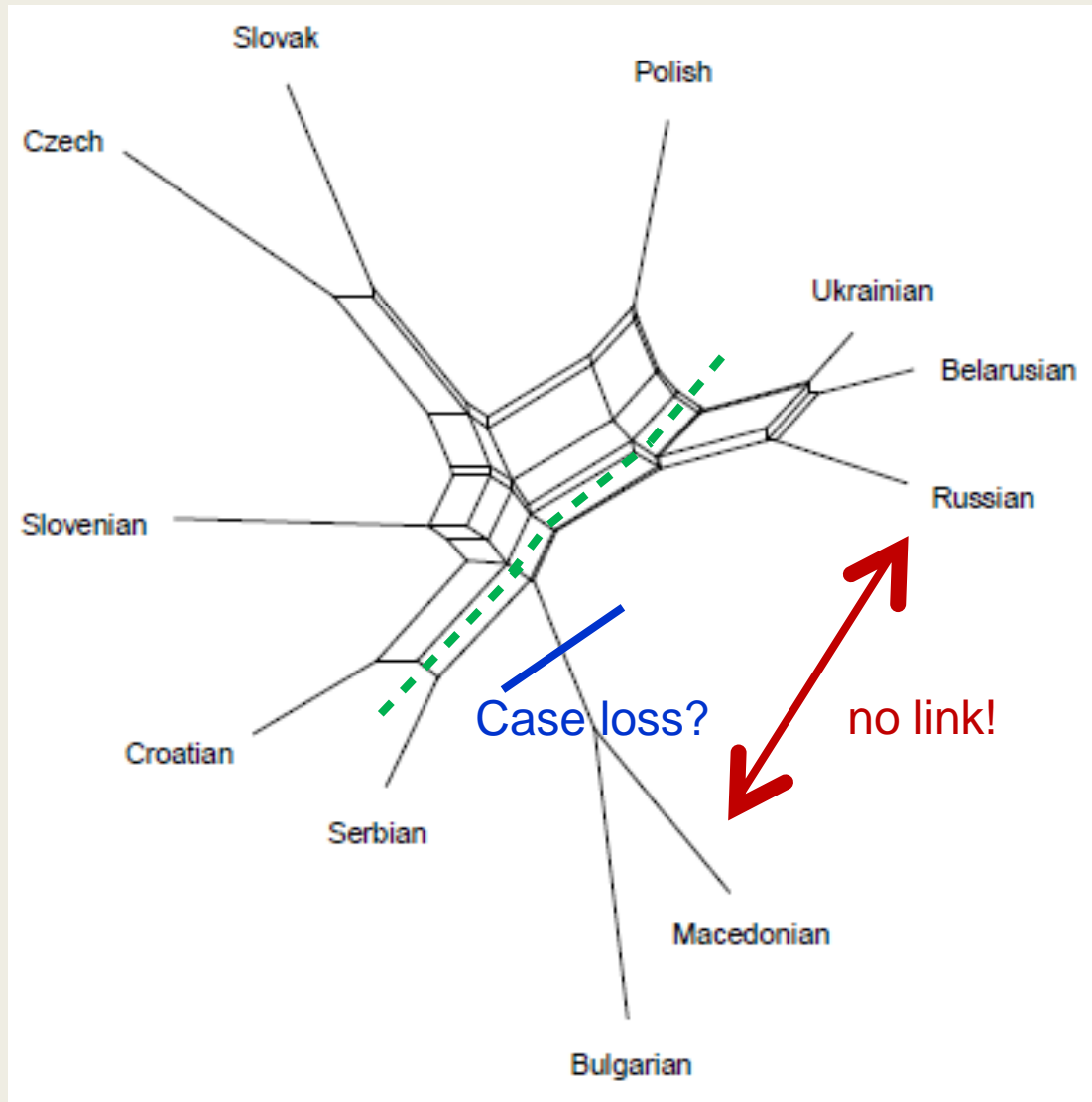
Areality



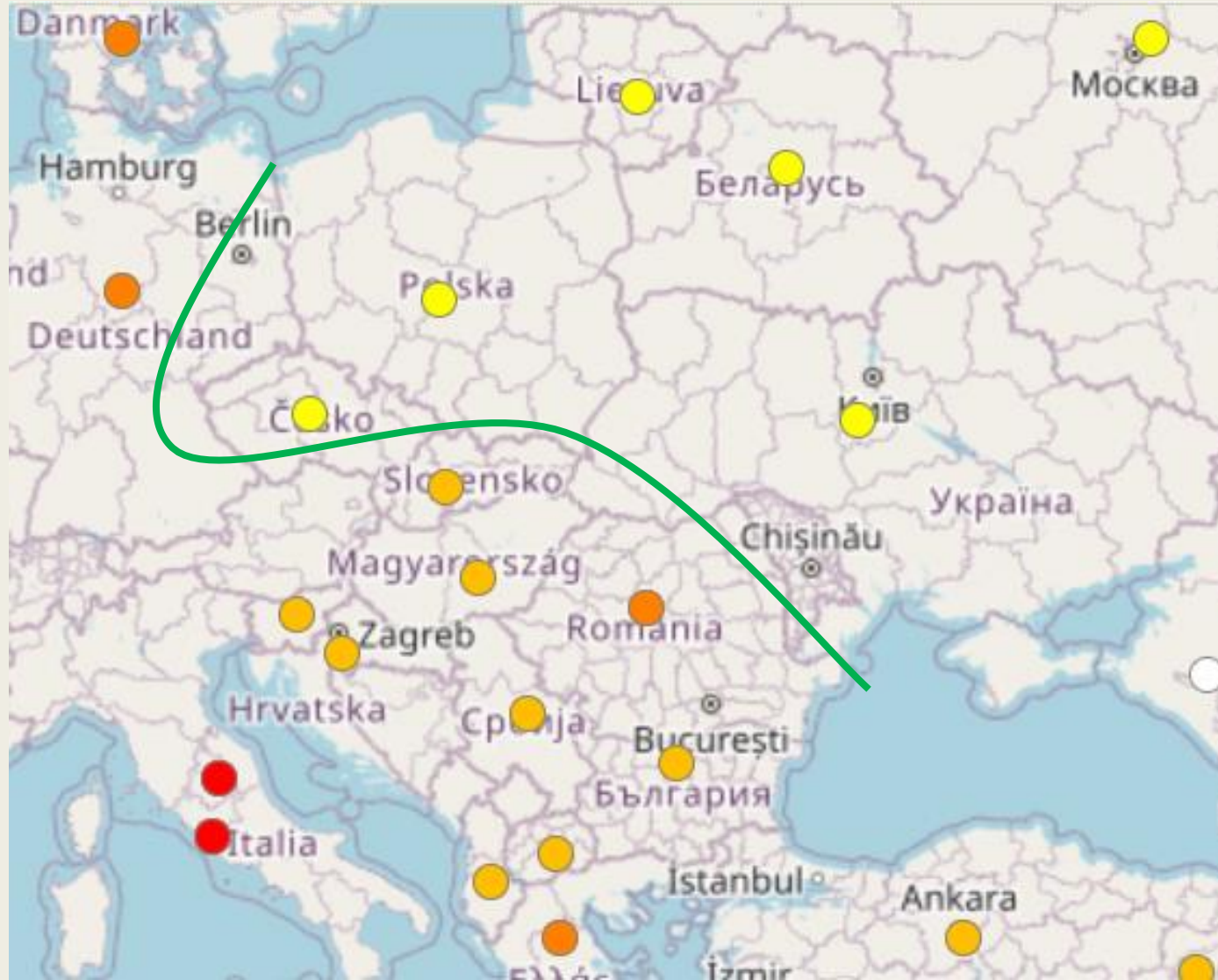
# Results: DistEtymVerb



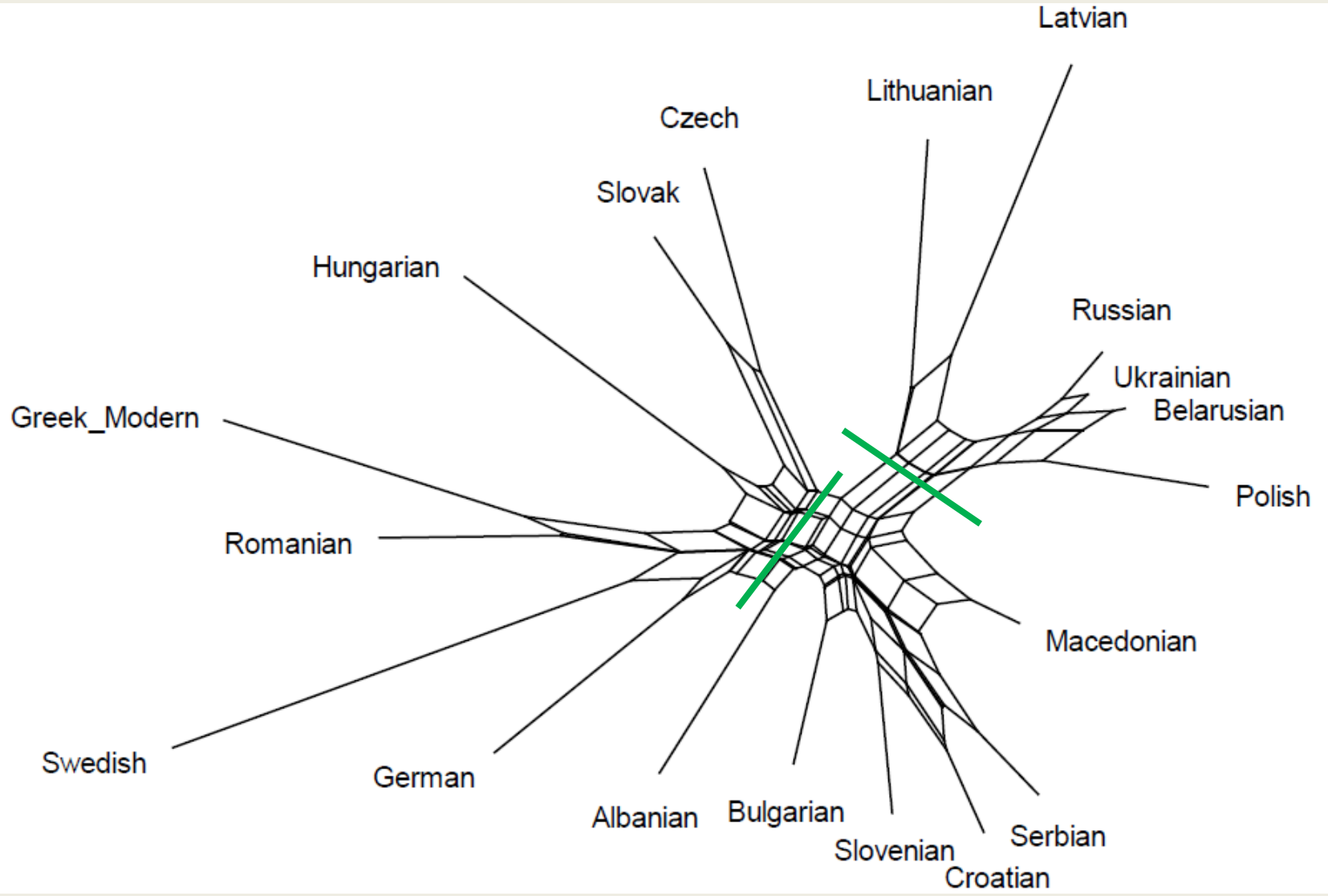
# Results: DistEtymPat



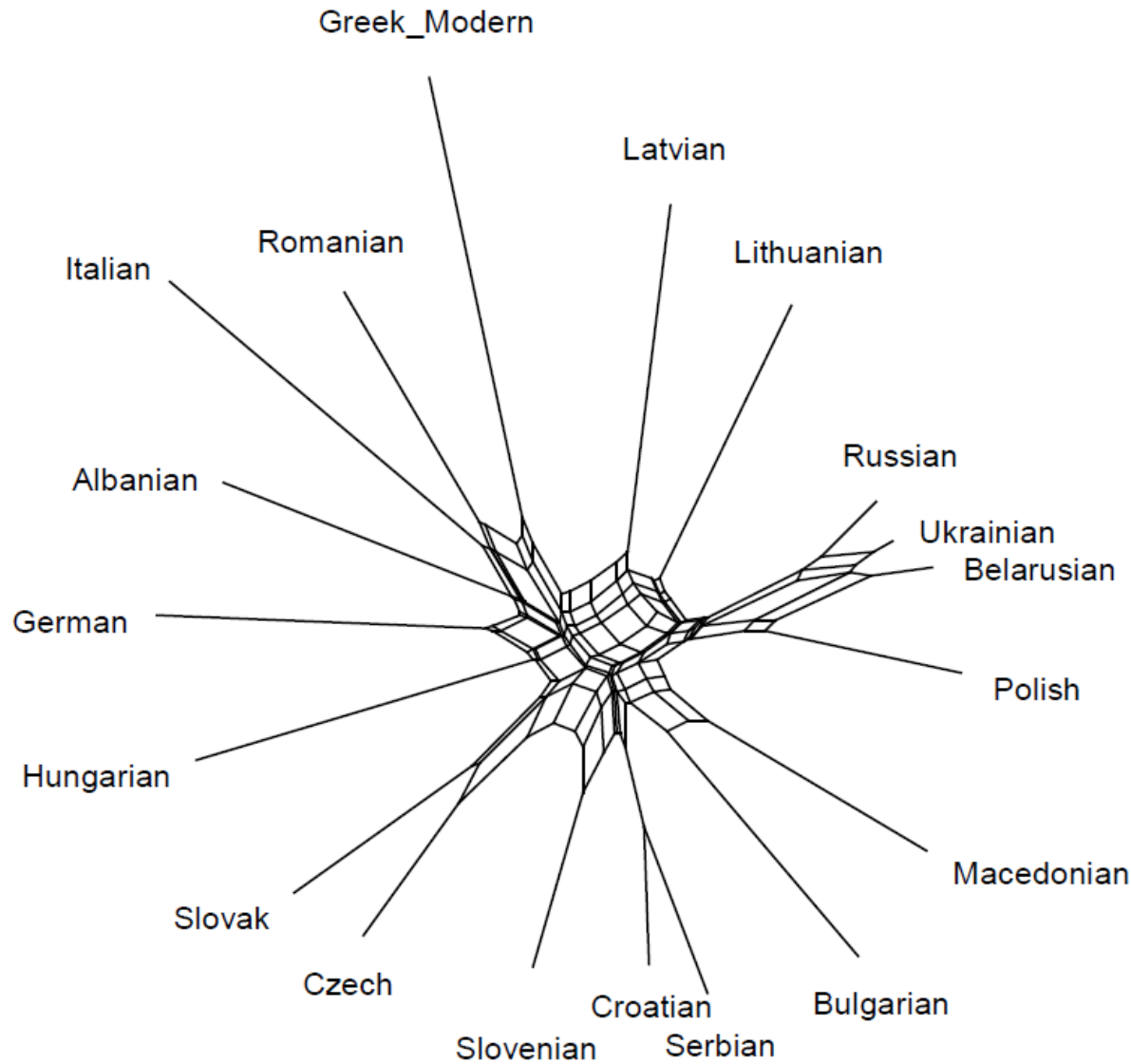
# Results: Transitivity Prominence



# Results: DistValLoc



# Results: DistValPat



Low-level signal

NB: PL vs. CZ+SK

# Summary

- Quantitative methods for assessing (dis)similarities
- For all phenomena, the areal dimension is not less important than the genealogical dimension, cf. systematic links between PL and UK+BR
- Links between Russian and BG/MK are visible in the verbs' cognacy relations, but not so much in valency patterns =>
  - *second South Slavic influence?*
  - *areal effects in genuine language contact (pattern borrowing) vs. cultural influences?*

# Summary

- Case loss in BG and MK affects cognacy relationships between specific valency encoding devices, but not so much the lexical extent of specific classes
  - *the renewal of valency classes is relatively independent of the renewal of encoding devices*
- “Deeper” valency-related phenomena, such as transitivity, locus and lexical composition of valency classes display wide-scope areal effects, whereby Slavic languages are part of the broader European landscape



THANK YOU!