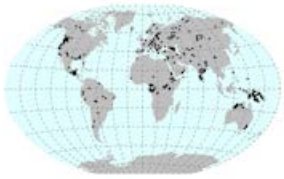
University of Bern
Department of Linguistics, ISW Colloquium
March 31, 2025

Genealogical and areal patterns in the distribution of bivalent valency class systems

Sergey Say
sergey.say@uni-potsdam.de
University of Potsdam

Universität Potsdam



PoSla Typology Lab

SFB 1287 Limits of Variability in Language
Cognitive, Computational, and Grammatical Aspects

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287

Structure of the talk

- **Setting the stage: typological study of valency**
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

Setting the stage

Prominent role of valency for linguistic typology

- transitivity
- alignment
- voice and related phenomena: passive, reflexive, …
- valency orientation

Setting the stage

Typology is mainly focused on major clause types

- monovalent: ‘sleep’, ‘run’, …
- transitive: ‘kill’, ‘break’, …
- ditransitive: ‘give’, …

Setting the stage

- All (?) languages have minor (a.k.a. non-canonical) valency patterns
- (Until recently) underrepresented in typological research

  «The selection principles apparently only govern argument selection for two-place predicates having a subject and a true direct object» [Dowty 1991: 576]

- Goal: to fill this gap for bivalent verbs

Setting the stage

- Why bivalent verbs?
  - they are especially prone to show deviant valency behaviour (Bickel et al. 2014)

(1) *The boy looked **at the clouds***

(2) *Das Heu duftet **nach** Pferd*

Russian

(3) ***Mne** nravitsja eta rubaška* ‘I like this shirt’

I.DAT like.PRS.3SG this.F.NOM.SG shirt.NOM.SG

  - they often form relatively large classes, unlike non-canonical trivalent verbs

Project at large: goals

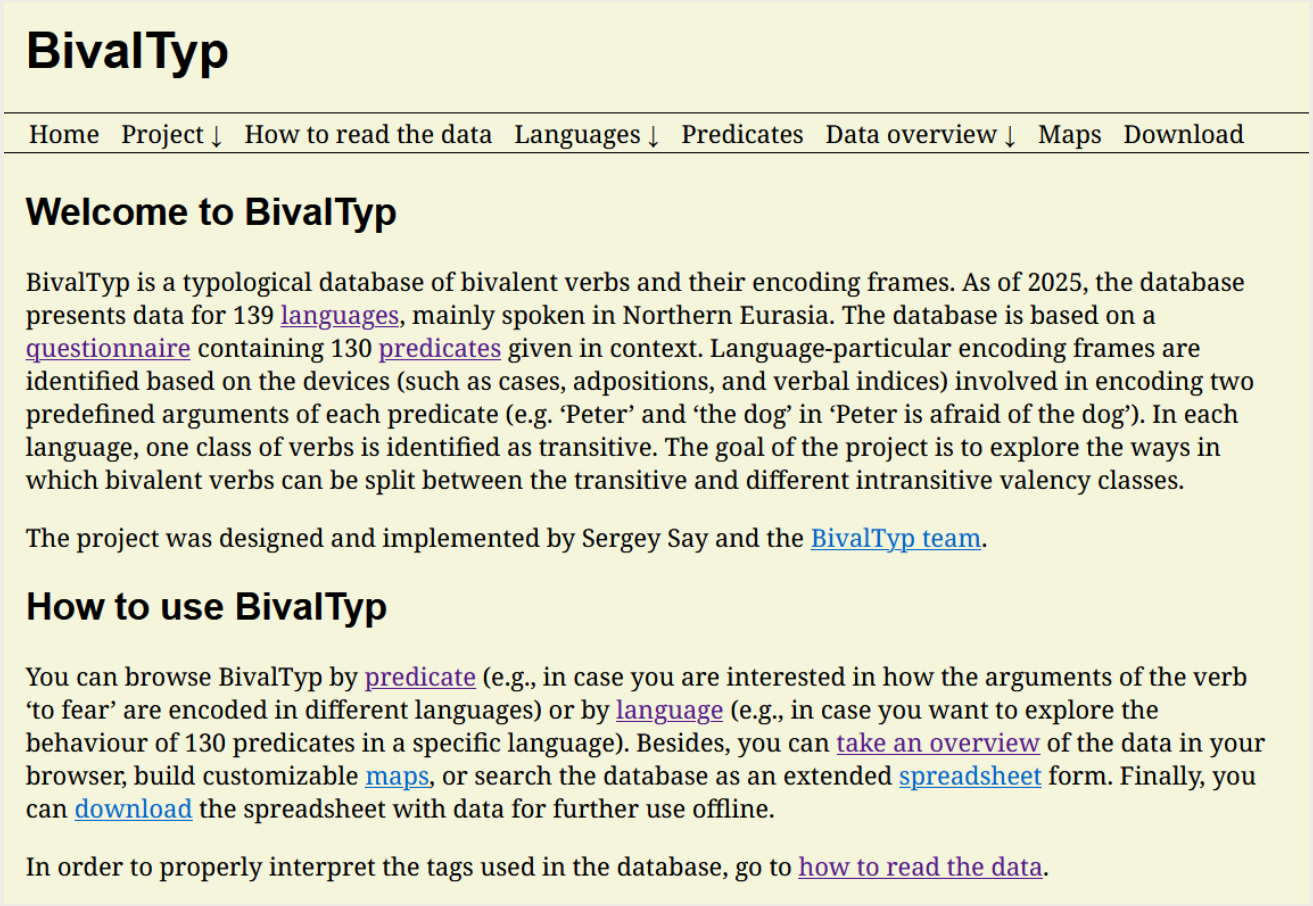
- Which factors determine valency class assignment in individual languages?
- How can we measure cross-linguistic differences in valency class systems?
- To what extent are valency classes (dis)similar in areally/genealogically related languages?
  - What is the depth of genetic effects = **how stable** are valency class systems?
  - What is the **granularity** of areal effects?

Structure of the talk

- Setting the stage: typological study of valency
- **The database: BivalTyp**
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

BivalTyp

Say, Sergey (ed.). 2020-. BivalTyp: Typological database of bivalent verbs and their encoding frames. (Available online at 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 2025, the database presents data for 139 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.

The project was designed and implemented by Sergey Say and the BivalTyp team.

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.

BivalTyp: Questionnaire

- First-hand data provided by language experts

- 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.

BivalTyp: X and Y

#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

⇒ Based on Dowty’s “lexical entailments” (1991), where “X” accumulates more agentive properties than “Y”

⇒ Identification of X and Y is independent of morphosyntactic coding in individual languages

BivalTyp: Valency pattern

- The valency of a verb = “the list of its arguments with their coding properties”
- Coding properties
  - flagging (cases & adpositions)
  - indexing (agreement, cross-referencing)
  - word order (rarely)

[*Den*	*Kindern*]	*gefällt*	[*der*	*Schneemann*].
the.PL.DAT	child.PL.DAT	please.3SG	the.SG.NOM	snowman.SG.NOM
‘The children like the snowman.’				

Malchukov and the Leipzig Valency Classes Project team, 2015: 30

BivalTyp: Valency patterns

- Each construction is annotated for its (language-specific) valency pattern: encoding of X and Y

- Coding devices: flagging (trivial), indexing (4), and word order (rarely)

Abaza (< Northwest Caucasian)

(4) *fatíma* *murád* *jə-z-qá-l-ç-əj-t*
PN PN [3SG.M.IO-BEN]-LOC-[3SG.F.ERG]-believe-PRS-DCL
'Fatima trusts Murad.'

=> Valency pattern = "ERG_BEN"

BivalTyp: Valency classes

- Two verbs belong to the same **valency class** iff their arguments are coded by the same combination of argument-encoding devices
  - E.g., in German #28 *warten* ‘wait’, #95 *schauen* ‘look’, #101 *schießen* ‘shoot’, and #122 *sauer sein* ‘have a grudge’ all belong to the [NOM_aufACC] class.

- Cross-linguistic variation in the number of attested classes
  - Min: 6 in Joola-Fonyi (Niger-Congo)
  - Max: 37 in Adyghe (Northwest Caucasian)

BivalTyp: Transitivity

- A pattern is considered **transitive** if its two arguments are morphosyntactially coded like the ‘killer’ and the ‘victim’ micro-roles of the verb ‘to kill’
  cf. Haspelmath (2015: 136)

- Coding devices involved in any language-specific transitive pattern are considered “core”; all other argument-coding devices are considered non-core

BivalTyp: Transitivity

- Russian

#105: *Пет-я* *уби-л* *Маш-у* => **TR** (by definition)
PN-NOM.SG kill.PST.M.SG PN-ACC.SG
'Petja killed Maša'

#28 *Пет-я* *ждет* *Маш-у* => **TR**
PN-NOM.SG wait.PRS.3SG PN-ACC.SG
'Petja is waiting for Maša'

#21 *Пет-я* *достиг* *берег-а* => **NOM_GEN**
PN-NOM.SG reach.PST.M.SG bank-GEN.SG
'Petja reached the bank'

BivalTyp: Locus

- Non-transitive patterns are classified based on whether one or both of the two arguments, X and Y, are encoded as non-core argument NPs.
- Four-way classification of patterns: TR, X, Y, XY
- Justification
  - "Deviations" from Hopper and Thompson's (1980) transitivity prototype are usually encoded on the relevant constituent (Malchukov 2005, 2006)
  - A technique that allows to abstract from the language-specific details in the organization of case paradigms and alignment patterns => a comparative concept

BivalTyp: Locus of ‘dream (about)’

Croatian: TR

(5)

Per-o	*sanja*	*Marij-u*
PN-NOM.SG	see_in_dream.PRS.3SG	PN-ACC.SG

‘Pero dreams about Marija.’

Polish: X

(6)

Roman-owi	*śni*	*się*	*Matk-a*	*Bosk-a*
PN-DAT.SG	dream:IPFV.PRS.3SG	REFL	mother-NOM.SG	god’s-F.NOM.SG

‘Roman dreams about Virgin Mary.’

Slovak: Y

(7)

Peter	*sníva*	*o*	*Márii*
PN(M)[NOM.SG]	dream(IPFV).PRS.3SG	about	PN(F)-LOC.SG

‘Peter dreams about Maria.’

Czech: XY

(8)

Petr-ovi	*se*	*zdá*	*o*	*Michal-ovi*
PN(M)-DAT.SG	REFL.ACC	dream(IPFV).PRS.3SG	about	PN(M)-LOC.SG

‘Petr dreams about Michal.’

BivalTyp: Sample

- The sample: currently 139 languages, mainly spoken in Northern Eurasia

BivalTyp: Contributors

- A big **THANK YOU** to language experts / contributors

Indira Abdulaeva, Anna Alexandrova, Daria Alfimova, Mansour Amadeh, Ekaterina Aplonova, Peter Arkadiev, Gilles Authier, David Avellan-Hultman, Aleksandra Azargaeva, Ayten Babaliyeva, Mislav Benić, Sandra Birzer, Alena Blinova, Natalia Bogomolova, Nadezhda Bulatova, Yura Chernov, Denis Creissels, Michael Daniel, Marah Deeb, Varvara Diveeva, Sergey Dmitrenko, Rhenee Espayos, Vladimir Fedorov, Timothy Feist, Tagir Gadzhiakhmedov, Martin Gális, Dmitry Ganenkov, Rowena Garcia, Uzlipat Gasanova, Dmitry Gerasimov, Wakweya Gobena, Elena Gorbova, Irene Gorbunova, Olga Gorickaja, Mariza Ibragimova, Ingunn Hreinberg Indriðadóttir, Ildar Ibragimov, Emil Ingelsten, Sylvanus Job, Vasilisa Kagirova, Maxim Kloczenko, Maria Khachaturyan, Rashidat Khalidova Maria Khazhomia, Maria Kholodilova, Mikhail Knyazev, Elena Kolpachkova, Daria (Suetina) Konior, Maria Konoshenko, Yukari Konuma, Elena Kordi, Richard Kowalik, Kirill Kozhanov, Irina Külmoja, Samona Kurilova, Nikita Kuzin, Olga Kuznecova, Natalia Logvinova, Nadege Nkwenti Lum, Timur Maisak, Anastasia (Borisovna) Makarova, Anastasia (Leonidovna) Makarova, Ramazan Mamedshaxov, Solmaz Merdanova, Stepan Mikhajlov, Daria Mischenko, Zarina Molochieva, George Moroz, Bulbul Musaeva, Majsarat Musaeva, Rasul Mutalov, Galina Nekrasova, Johanna Nichols, Dmitry Nikolaev, Ajtalina Nogovitsyna, Kwadwo Brobbey Nti, Sofia Oskolskaya, Maria Ovsjannikova, Anastasia Panova, Elena Perekhvalskaja, Natalia Perkova, Krasimira Petrova, Inna Popova, Maria Pupynina, Tatiana Repnina, Monika Rind-Pawlowski, Neige Rochant, Alexander Rostovtsev-Popiel, Daria Ryzhova, Ugur Sermiyan, Sergey Say, Ekaterina Sergeeva, Ksenia Shagal, Mayya Shlyakhter, Natalia Stoynova, Ksenia Studenikina, Gasangusen Sulaibanov, Nina Sumbatova, Evgenija Teplukhina, Mladen Uhlik, Naida Vagizieva, Anastasia Vasilisina, Arseniy Vydrin, Valentin Vydrin, Alena Witzlack-Makarevich, Elizaveta Zabelina, Natalia Zaika, Andreja Žele, Ekaterina Zheltova, Vasilisa Zhigulskaja, Daria Zhornik, Anastasia Zhuk`

BivalTyp: Entry structure

Telugu

54. fill (intr) *(niṁḍipōvu)*

Valency pattern: NOM_COM
X: NOM
Y: COM
Locus: Y

bākeṭ	*niḷḷa-tō*	*niṁḍipōyiṁdi*
bucket(N).SG.NOM	water.PL.OBL-COM	fill_up.PST.3SG.NM
‘The bucket filled with water.’		

55. find *(doruku)*

Valency pattern: DAT_NOM
X: DAT
Y: NOM
Locus: X

pravīṇ-ki	*tana*	*tāḷaṁcēvu-lu*	*dorikāyi*
PN(M).SG.OBL-DAT	own	key(N)-PL.NOM	be_found.PST.3PL.N
‘Praveen found his keys.’			

BivalTyp: Dataset

- 16574 entries (130 predicates in 139 lgs – 1496 gaps):
  - language ID
  - predicate ID
  - verb
  - valency pattern
  - (for 97 languages: interlinearized examples)
- The database is searchable, sortable and mappable by predicates, languages, valency patterns, etc.

- **Further contributions are very welcome!**

BivalTyp: possible applications

- transitivity ratio of verbs
- (dis)similarity between verbs
- predictability of valency patterns
- typologically informed analysis of language-specific valency class systems
- **(dis)similarity between languages**
- **comparison with genealogical and areal data**
- comparison with structural data: case, WO, etc.
- and many more

Structure of the talk

- Setting the stage: typological study of valency
- The database: BivalTyp
- **Cross-linguistic comparison**
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

Cross-linguistic differences: Casual intro

Table 1. Selected verbs and valency patterns in Joola-Fonyi and Khwarshi.

	Joola-Fonyi		Khwarshi	
meaning	verb	pattern	verb	pattern
'be afraid'	*kólí*	TR	*j/uƛ'a*	ABS_CONT
'avoid'	*ŋom*	TR	*j/ič̣a*	ABS_CONT.EL
'wait'	*kob*	TR	*gic'a*	ABS_CONT.LAT
'attack'	*lóúm*	TR	*k'oƛa*	ABS_SUPER
'win, beat'	*ŋoolen*	TR	*j/iža*	ABS_SUPER.EL
'see'	*juk*	TR	*j/akʷa*	DAT_ABS
'touch'	*gor*	TR	*j/etaχa*	ERG_CONT
'bite'	*rum*	TR	*hana*	ERG_GEN1
'be angry'	*leet*	TR	*semi mak'a*	GEN1_CONT.LAT
'eat'	*ri*	TR	*j/ac'a*	TR (ERG_ABS)

Cross-linguistic comparison:
The scale of geographical patterning

- "The scale of geographical patterning is the size of the areal unit – local, subcontinental, larger than continental, global – within which the geographical distribution of a feature displays some clear and describable pattern. For example, … nominal classes tend to cluster areally and form hotbeds which are generally smaller than continental in size (subcontinental)"
  (Nichols 1992: 185)

Cross-linguistic comparison: Overview

- Genealogical dimension
- Geographic dimension
- Structural dimensions: various aspects of valency class systems

Genealogical dimension

- Three levels, based on WALS
  - 1: same genus
  - 2: same family, different genera
  - 3: different families

E.g.: DistGen (Eastern Armenian, Azerbaijani) = 3

Geographic dimension

- Geographic distance: calculated as the geographic distance (in kilometers) between the two points associated with individual languages
- Coordinates are mainly taken from Glottolog
- The distance is calculated using `distCosine()` from the R package `geosphere` (Hijmans 2016)
- NB: this is a very coarse metric for languages spoken over vast areas
- For statistical purposes, the decimal logarithm of the distance is used, e.g.

DistGeo (Eastern Armenian, Azerbaijani) = 277 km

LogDistGeo (Eastern Armenian, Azerbaijani) = 2.44

Structural dimensions

- Valency patterns cognacy: DistValEtym
- Transitivity Prominence
- Lexical transitivity profiles: DistValTrans
- Lexical locus-based profiles: DistValLoc
- Lexical distributions into language-specific valency classes: DistValPat

Structure of the talk

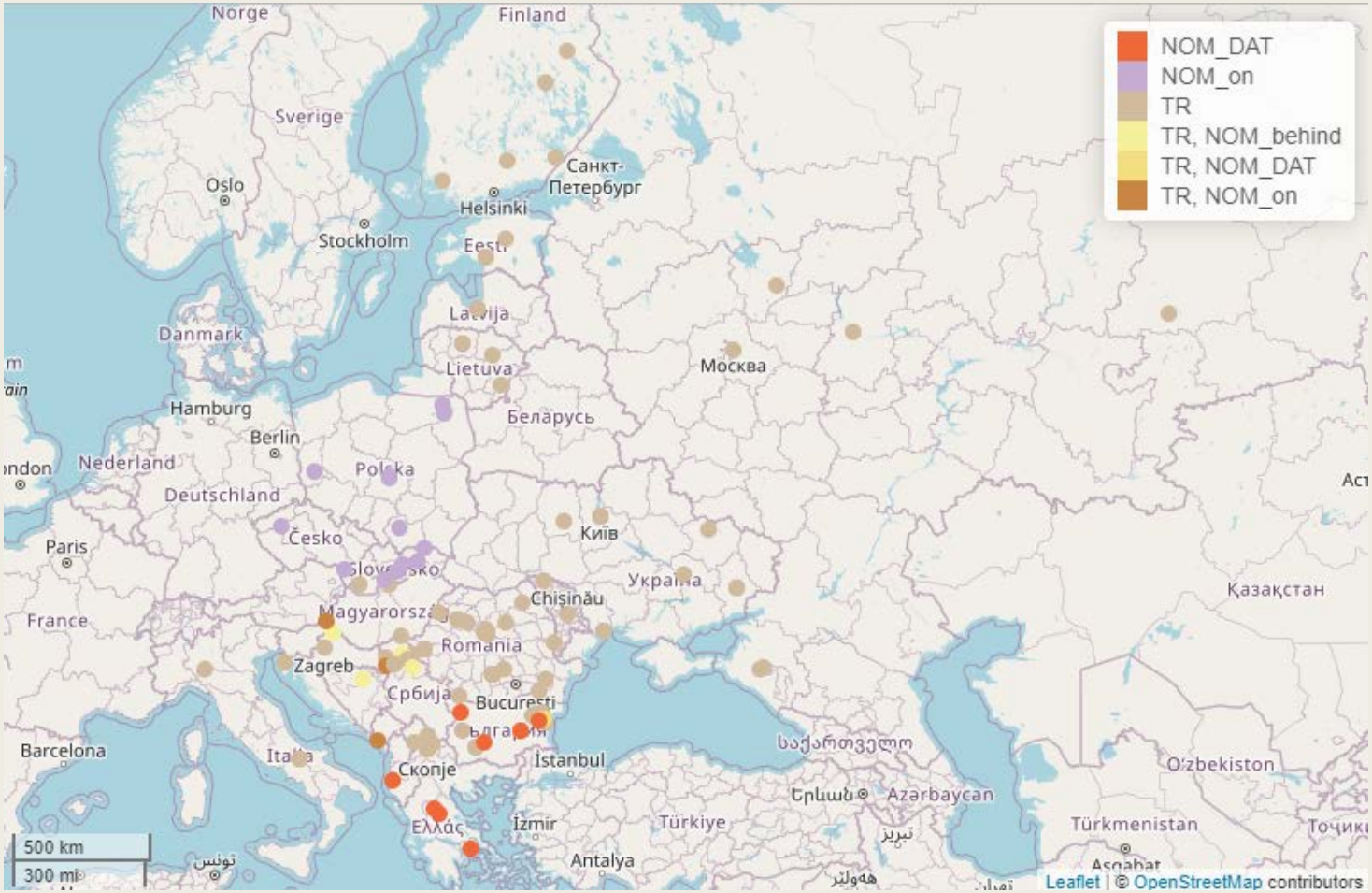
- Setting the stage: typological study of valency
- The database: BivalTyp
- Cross-linguistic comparison
  - **Valency patterns cognacy**
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

Valency patterns cognacy

- The logic: valency patterns of a given predicate in two varieties are considered “the same” if they involve cognate argument-coding devices

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

‘wait’ in Romani dialects



Kirill Kozhanov, Sergey Say. 56th Annual Meeting of the Societas Linguistica Europaea. August 29 – September 1, 2023. National and Kapodistrian University of Athens. Genealogy vs. contact configuration: argument coding across Romani dialects in Europe.

Patterns cognacy: DistValEtym

- Convert raw data into a distance metric
- The relative Hamming distance: the ratio of predicates that are assigned to different valency classes in two languages

Patterns cognacy: DistValEtym

	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, DistValEtym (UK, PL) = 0.203 = 26/128

Patterns cognacy: DistValEtym

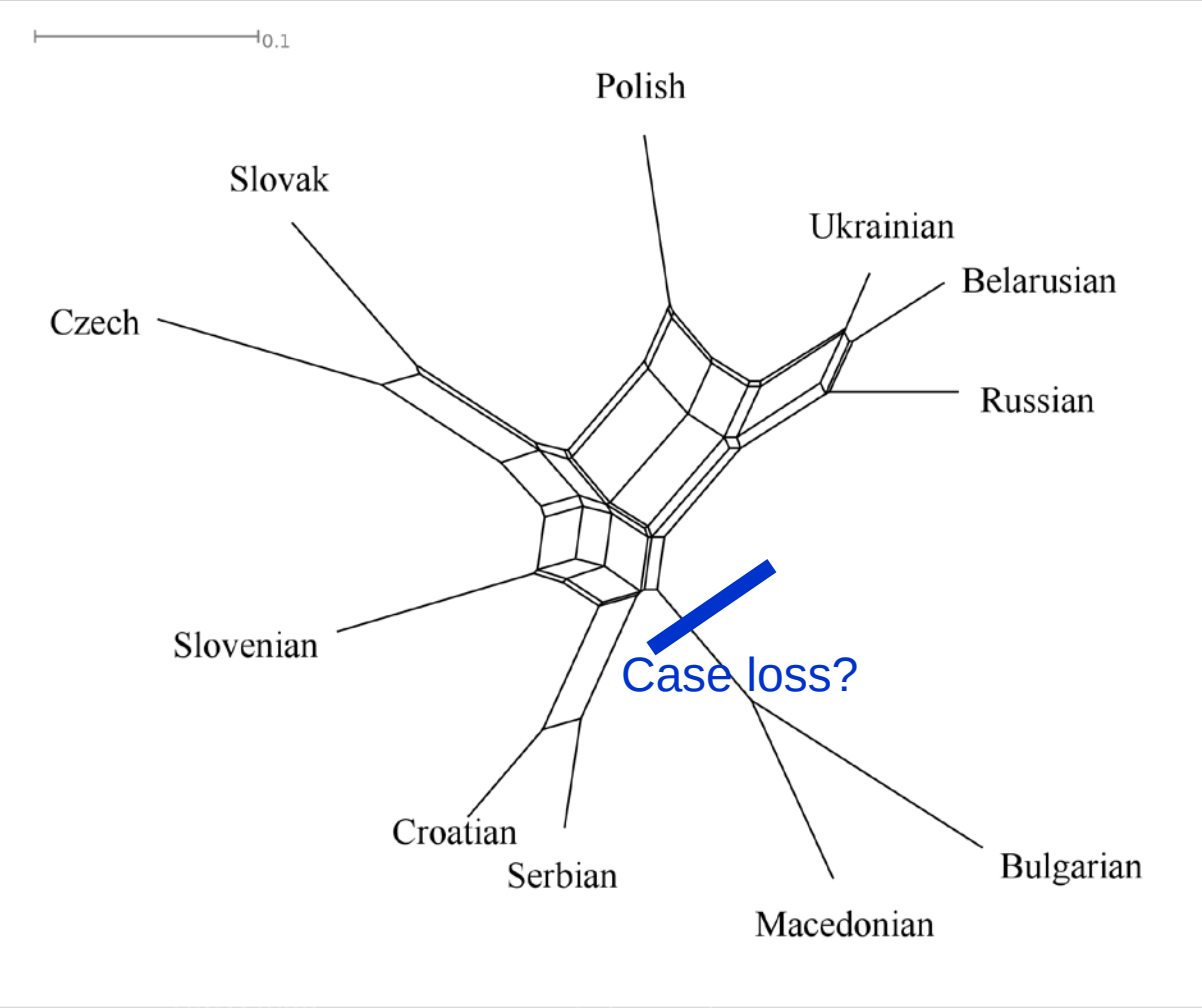
- Using DistValEtym allows to build a distance matrix: a matrix of distances in pairs of languages

	Russian	Polish	Serbian	Ukrainian	Slovenian	Czech	...
Russian	0,00	0,27	0,30	0,12	0,34	0,40	
Polish	0,27	0,00	0,35	0,20	0,34	0,36	
Serbian	0,30	0,35	0,00	0,30	0,25	0,40	
Ukrainian	0,12	0,20	0,30	0,00	0,34	0,40	
Slovenian	0,34	0,34	0,25	0,34	0,00	0,32	
Czech	0,40	0,36	0,40	0,40	0,32	0,00	
...							

Patterns cognacy: DistValEtym

- The usual problem: high dimensionality
- Standard algorithms
  - Multidimensional scaling
  - NeighborNet
  - Hierarchical clusterization

DistValEtym: Slavic



Both areal and genealogical dimensions play a role, see the position of:

- Polish
- Slovenian

NeighborNet visualization of the DistValEtym distances, see also Seržant et al.' (2022: 323)

DistValEtym: Limitations

- Valency classes are language specific (Haspelmath 2009: 510; Comrie et al. 2015: 4–5)
  - 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-qraʔa-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	*меʜemed.i-qaj*	*uqː.a-a*
PN[ABS]	PN-COM	fight.IPF-PRS

'Aslan is fighting with Muhammad.' => ABS_COM

DistValEtym: Limitations

- Cognacy-based logic is applicable for dialects of the same language or closely related languages
- DistValEtym is **not** applicable to remotely related or unrelated languages
- **Transitivity**, defined as a comparative concept, is one possible *tertium comparationis*, see the next section

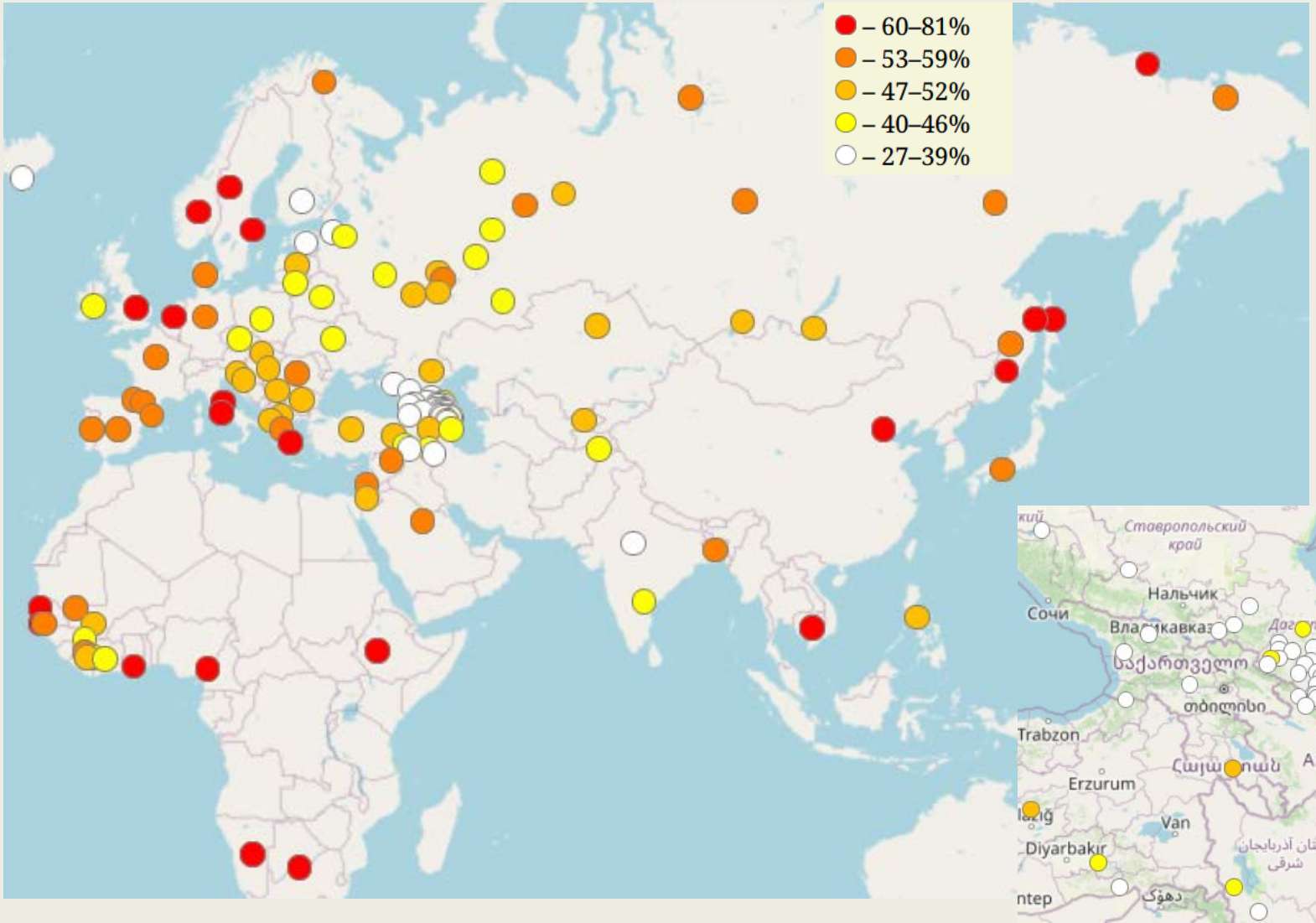
Structure of the talk

- Setting the stage: typological study of valency
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - **Transitivity prominence**
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

Transitivity prominence

- Transitivity prominence = the number of transitive entries divided by the total number of entries in the dataset (Haspelmath 2015)

Transitivity prominence: Areality



Transitivity prominence

- Transitivity prominence displays robust areal patterning
- High transitivity prominence values in Western and Southern Europe as well as in the Far East
- Low values are concentrated in Eastern Europe and particularly in the Caucasus, cf. observations scattered in the literature (Bossong 1998; Haspelmath 2015: 139–142; Lazard 1994: 63; Say 2014)

Transitivity profiles

- Setting the stage: typological study of valency
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - **Lexical transitivity profiles**
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- Conclusions

Transitivity profiles

- Transitivity prominence is a summary measure that disregards details

- When two languages have the same transitivity prominence score, they do not necessarily assign the same verbs to the transitive class

Transitivity profiles

	Eastern Armenian	Azerbaijani
win	TR	INTR
be_afraid	INTR	INTR
believe	INTR	INTR
see	TR	TR
reach	INTR	INTR
touch	INTR	INTR
forget	INTR	TR
wait	TR	TR
know	TR	TR
avoid	INTR	INTR
…		

- 2 mismatches out of 10 trials in this toy example

Transitivity profiles

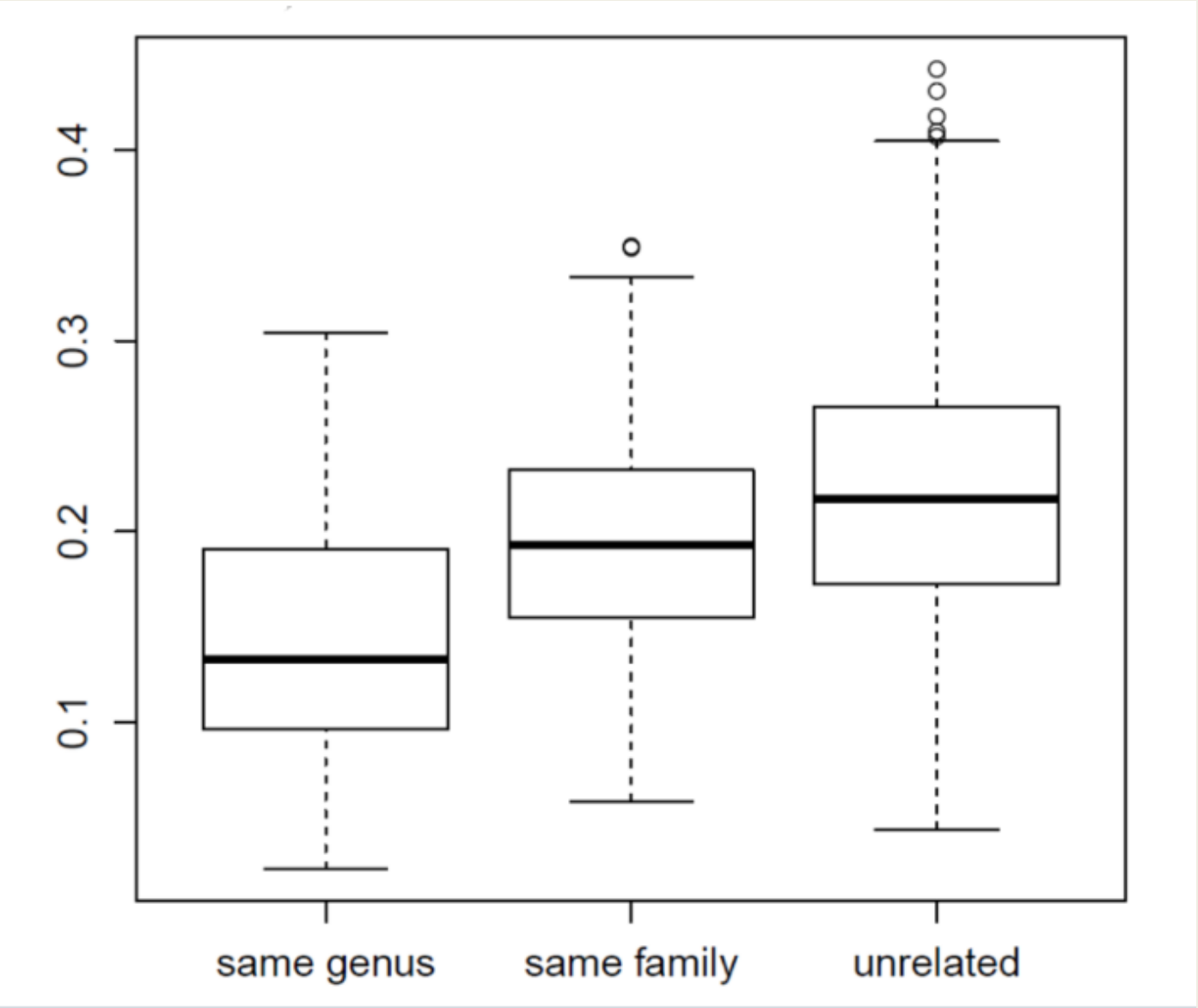
		Azerbaijani	
		TR	INTR
Eastern Armenian	TR	53	8
	INTR	5	53

DistrValTrans (Eastern Armenian, Azerbaijani) = (5+8)/(53+8+5+53) = 13 / 119 = 0.109

Transitivity profiles

- Significant genealogical signal at the level of individual genera, weaker signal at the family-size level

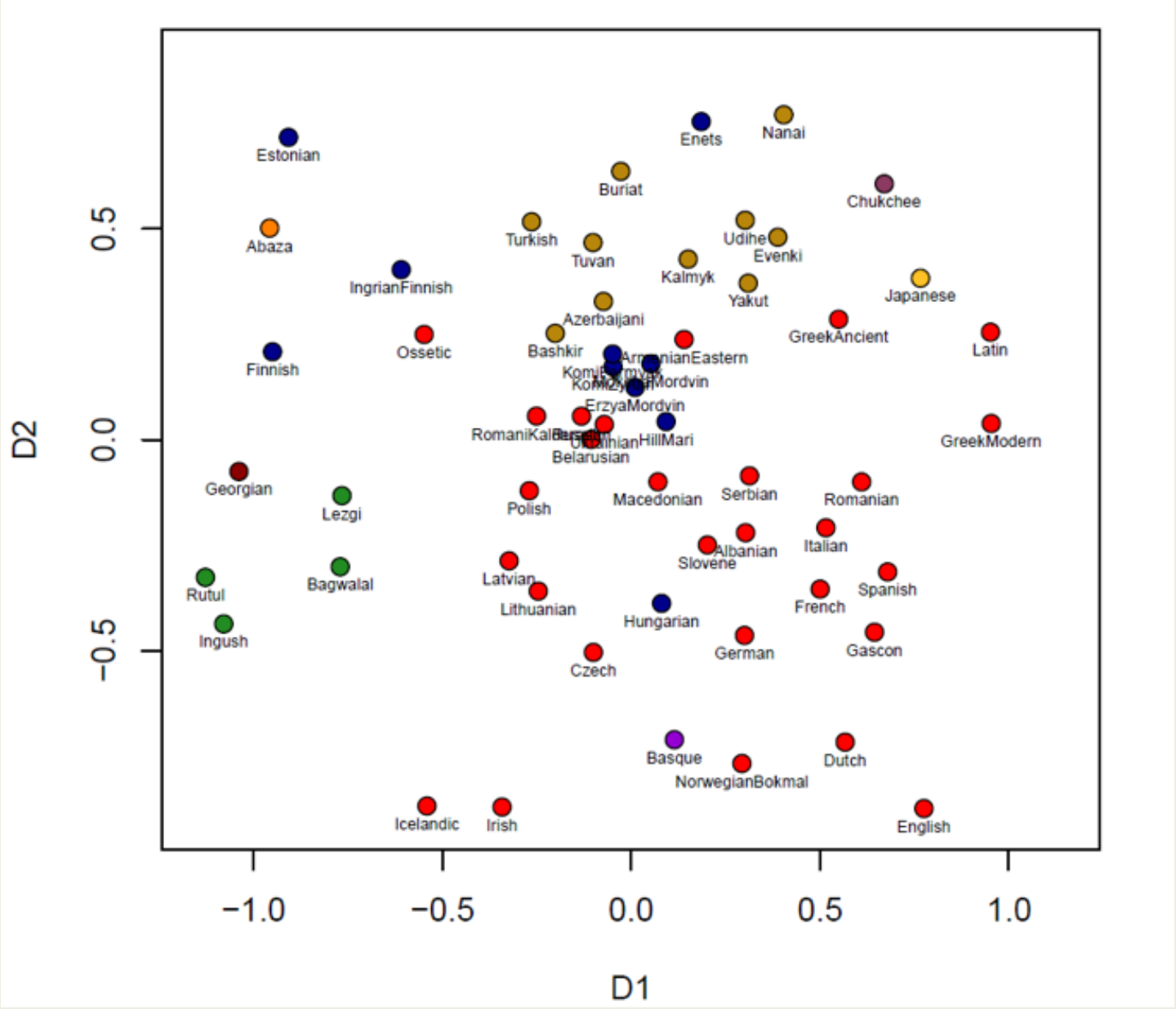
DistValTrans & genealogical distance



DistValTrans: Areal effects

- E.g., Uralic languages are distorted due to language contact:
  - Enets patterns with other languages of Siberia
  - Hungarian patterns with Standard Average European languages
  - Permic, Mordvinic and Mari are between Slavic and Altaic
  - Baltic Finnic languages are unlike anything else

DistValTrans: select languages (MDS)



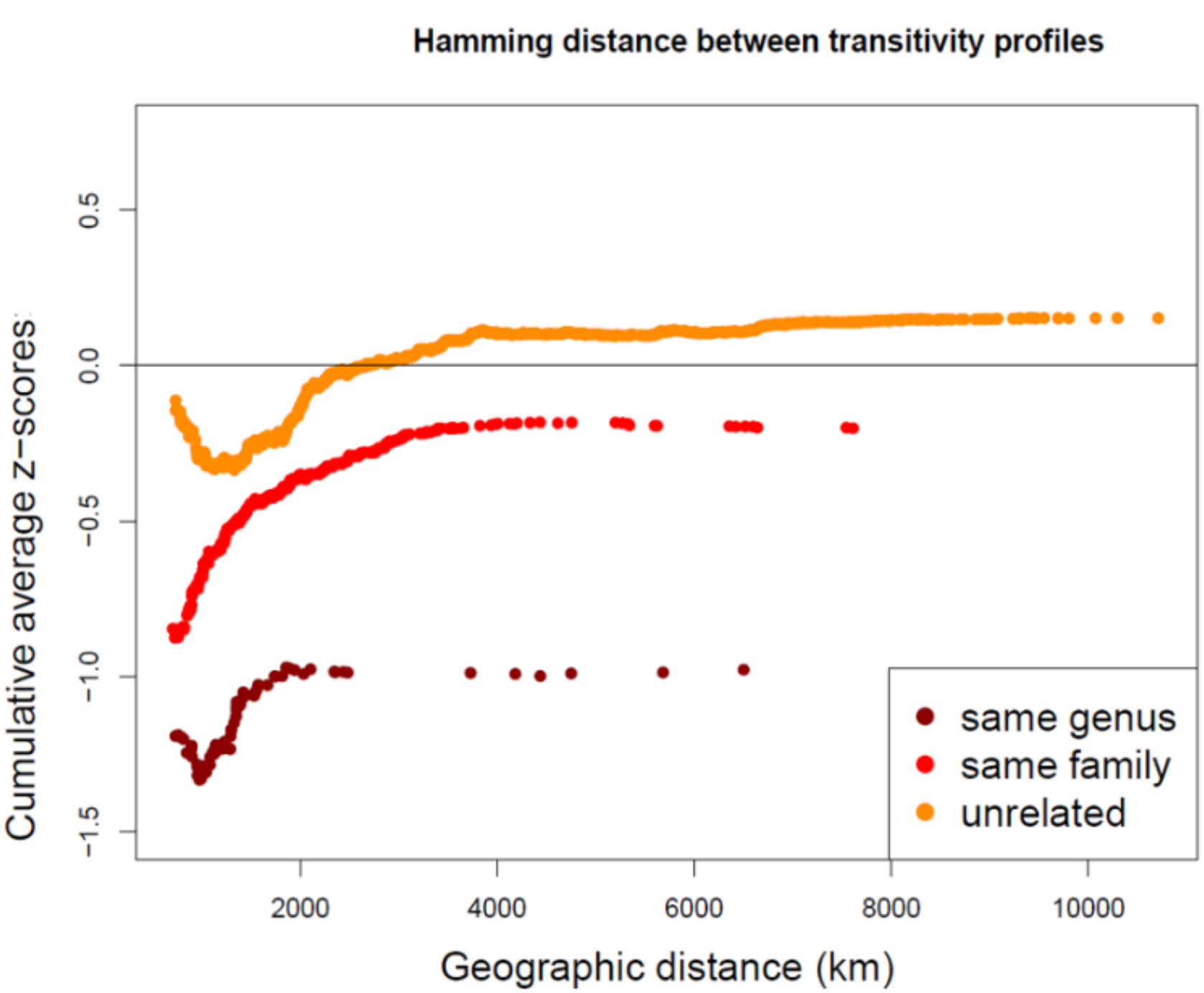
Transitivity profiles: Verb hierarchies

- Evidence for family-specific verb hierarchies of transitivity prominence
  - Experiential predicates (‘see’, ‘know’, ‘love’, ‘want’) are especially prone to be intransitive in Nakh-Daghestanian
  - Verbs of contact (‘follow’, ‘reach’, ‘touch’, ‘kiss’, ‘attack’) are especially prone to be intransitive in Uralic (though not Hungarian)
  - etc.

Transitivity profiles

- Next slide: the role of geographic distance
  - X-axis: geographic distance in kilometers
  - Y-axis: mean normalized DistValTrans values (z-scores) for pairs of languages spoken closer than N kilometers to each other (cumulative mean)
  - separately for three levels of genetic distance

This method is inspired by (Wichmann & Holman 2009: 75 ff.)


Hamming distance between transitivity profiles
Cumulative average z−scores
0.5
0.0
−0.5
−1.0
−1.5
2000
4000
6000
8000
10000
Geographic distance (km)
same genus
same family
unrelated

Transitivity profiles

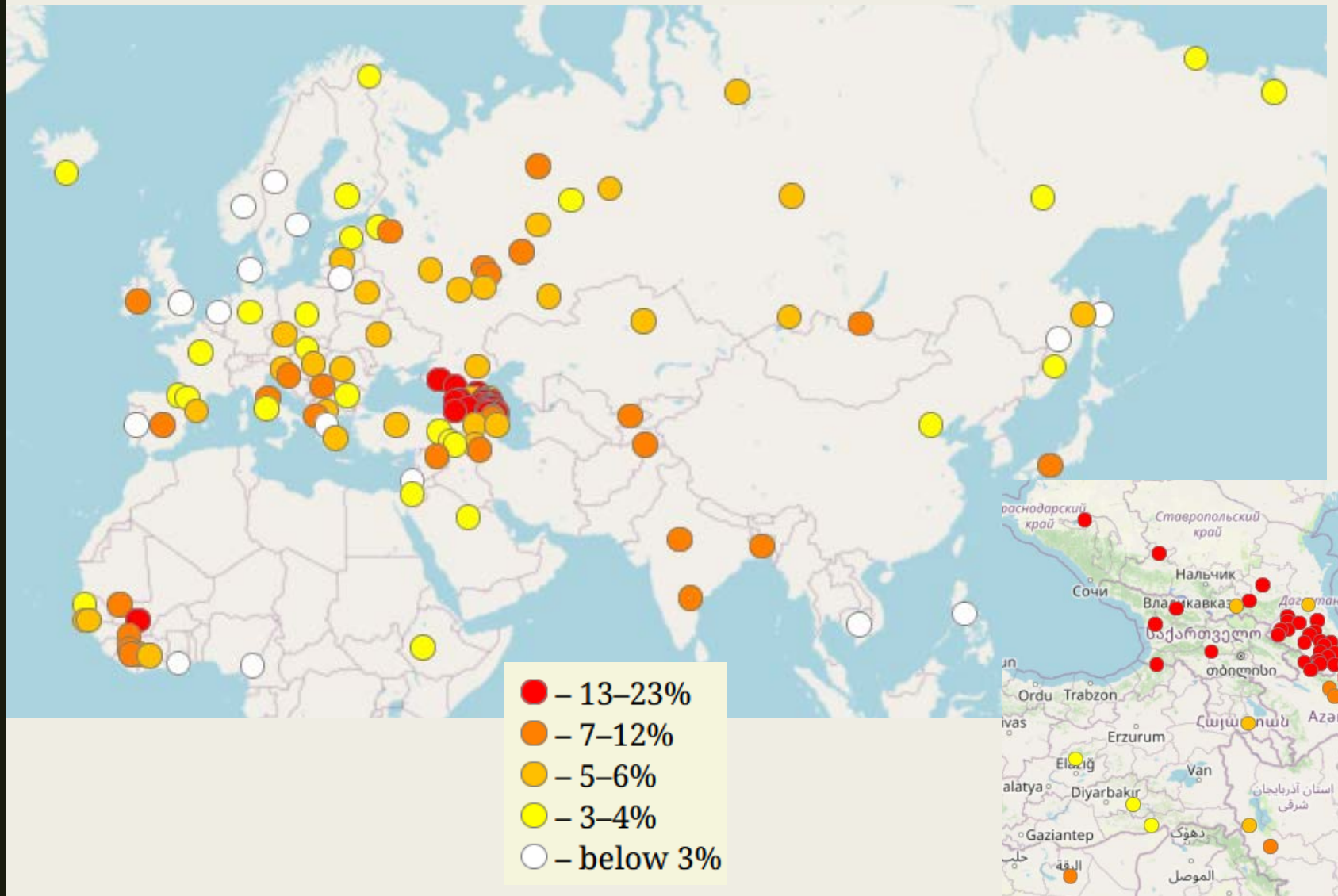
- Genealogical signal: three curves are very different

- If genealogical factor is levelled out, the role of geographic proximity rapidly fades away after ≈2000 km

Structure of the talk

- **Setting the stage: typological study of valency**
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - **Lexical locus-based profiles**
  - Lexical distributions into language-specific valency classes
- Conclusions

Locus-based differences: X-locus



Locus-based profiles: DistValLoc

- Same logic as in the case of DistValTrans, but with four possible loci as comparative concepts: TR, X, Y, XY

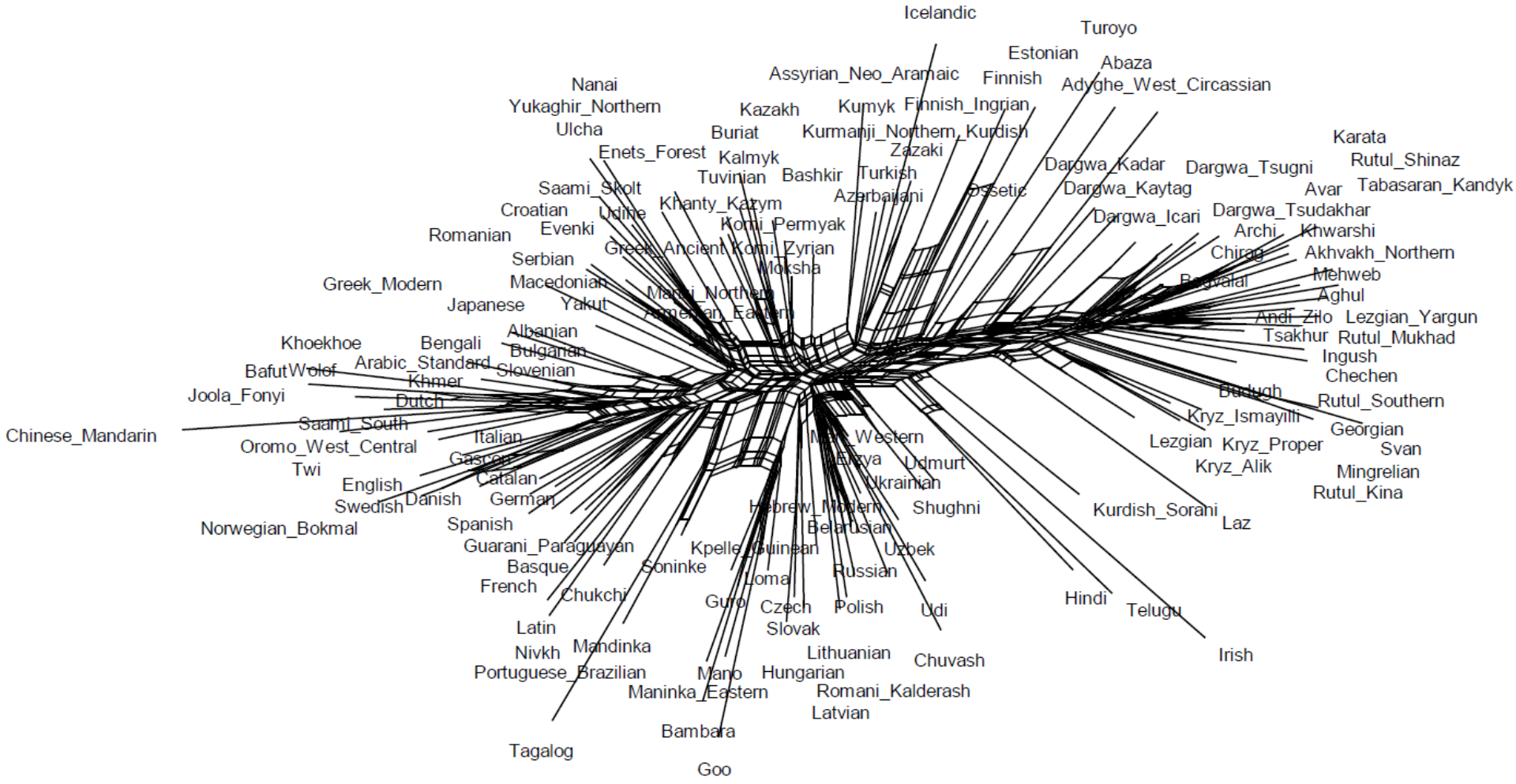
- DistValLoc > Distance matrix > Dimensionality reduction

Locus-based profiles: DistValLoc

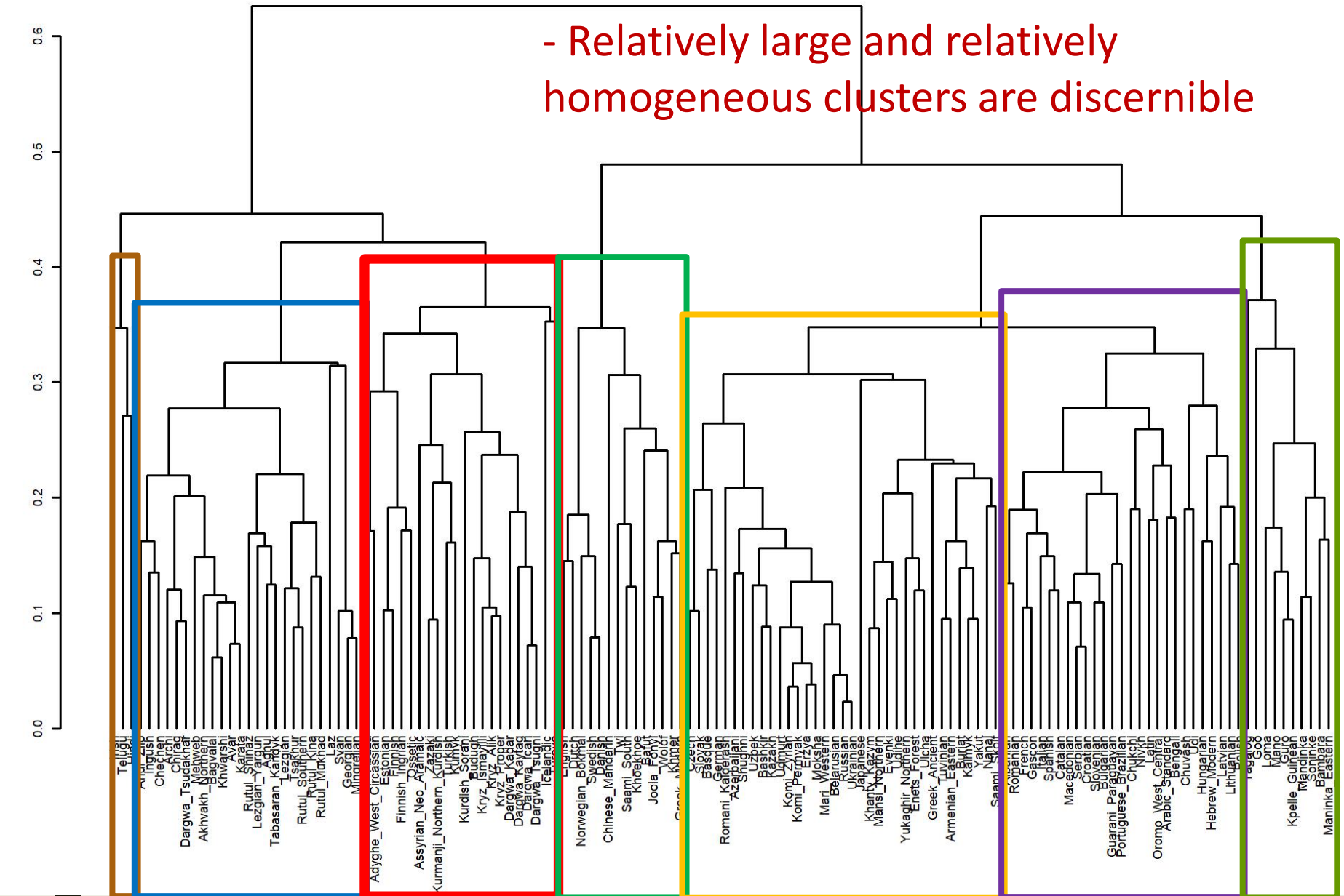
	Georgian	Turkish
forget	X	TR
call	Y	TR
know	Y	TR
avoid	Y	Y
make	TR	TR
have	X	X
look_for	Y	TR
bite	Y	TR
flatter	Y	TR
love1	X	TR
…		

- 7 mismatches out of 10 trials in this toy example

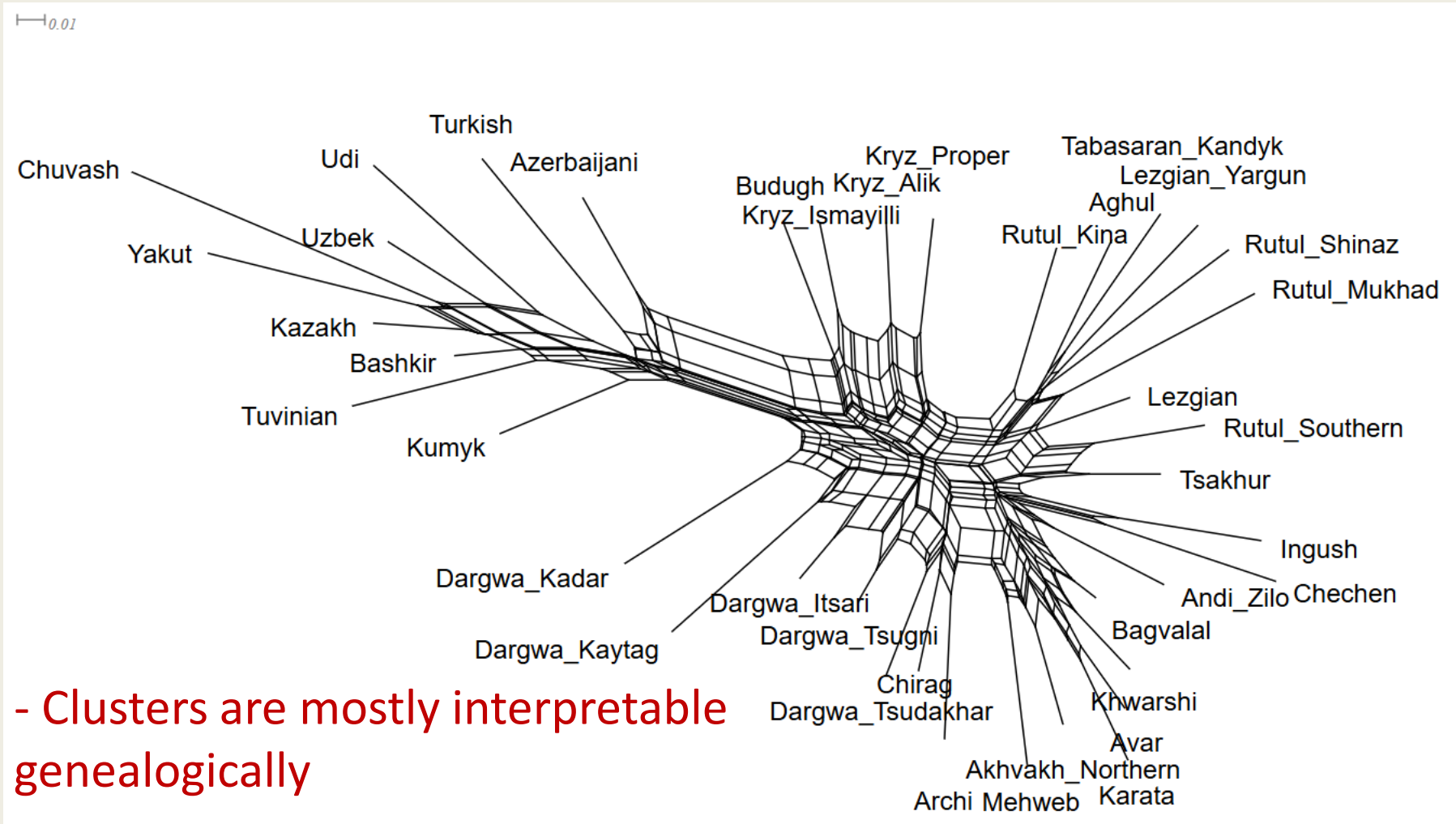
DistValLoc: Largely one-dimensional distribution



Hierarchical clusters based on DistValLoc



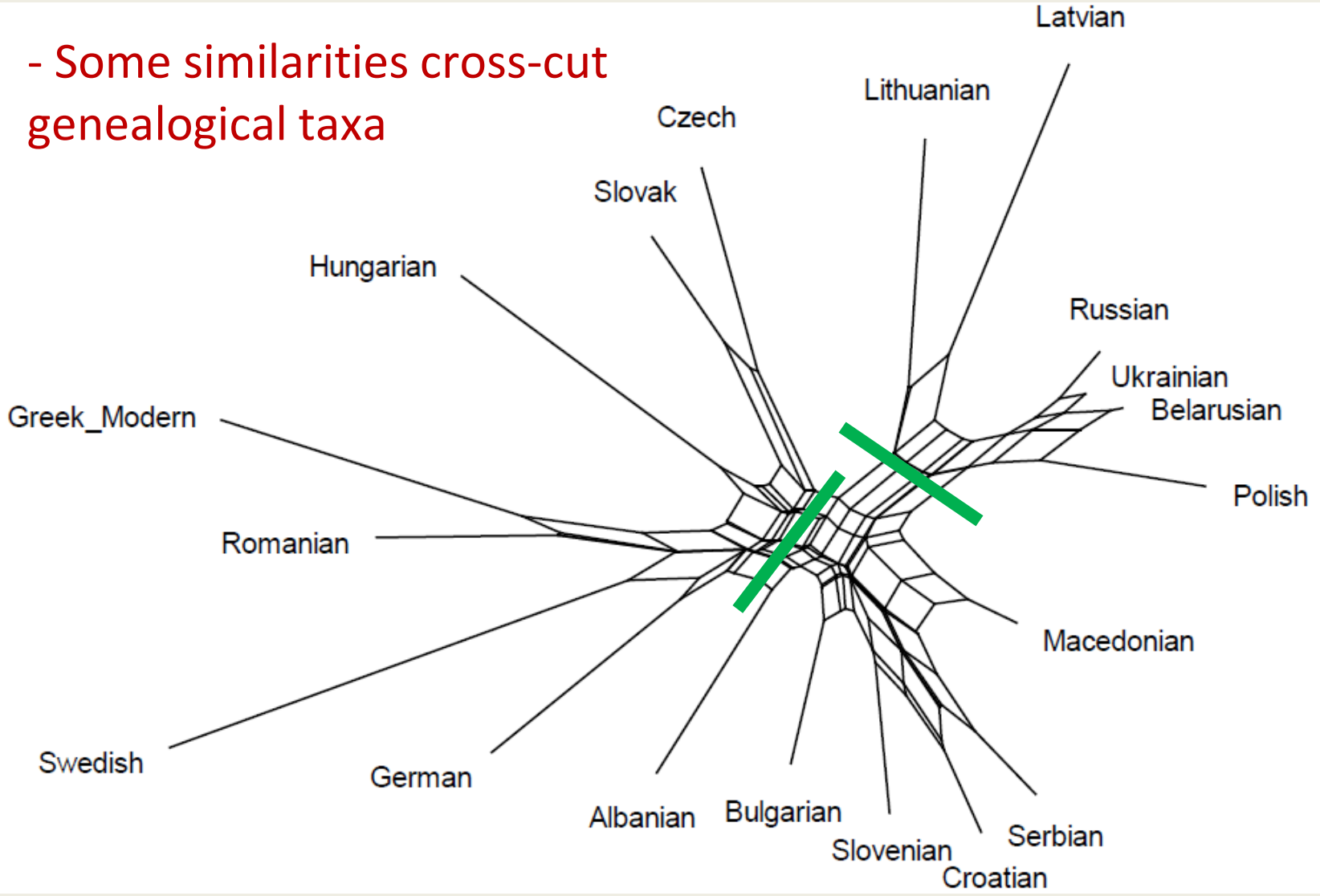
DistValLoc: Turkic + Nakh-Daghestanian



- Clusters are mostly interpretable genealogically

DistValLoc: Slavic + neighbors

- Some similarities cross-cut genealogical taxa



Structure of the talk

- Setting the stage: typological study of valency
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - **Lexical distributions into language-specific valency classes**
- Conclusions

Lexical distributions

- How can we compare valency class systems without
  - equating argument-coding devices (DistValEtym)
  - assuming gross types on *a priori* grounds (e.g. locus-based)?

- My solution: **DistValPat**, a metric based on entropy and MI (mutual information)

- Entropy ≈ the amount of information (conveyed by the valency class assignment)

Lexical distributions

- Wordlist-based typology
  - use the lexical **lists** as a *tertium comparationis* = set-partition variable
- Entropy: informal introduction
  - Shannon's entropy is a mathematical tool aimed at quantizing the amount of information associated with a certain discrete variable

Lexical distribution

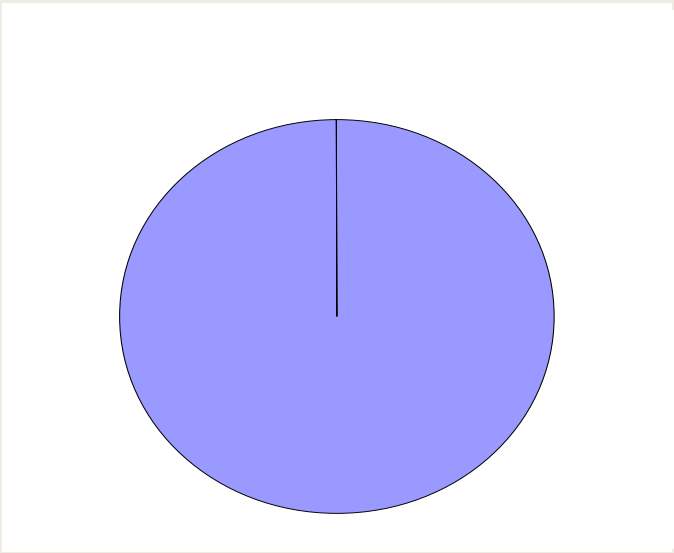
Eastern Armenian

#	Predicate	Translation	Valency Class
…			
21	reach	*Petros-ə hasav ap'-i-n* Petros[**NOM**]-DEF reach:AOR:**3SG** bank-**DAT**-DEF 'P. reached the bank'	**NOM_DAT**
22	touch	*Petros-ə dipav pat-i-n* Petros[**NOM**]-DEF touch:AOR:**3SG** wall-**DAT**-DEF 'Petros touched the wall'	**NOM_DAT**
53	attack	*Arǰ-ə harjakvec' jknors-i vra* bear[**NOM**]-DEF attack:AOR:**3SG** fisherman-**DAT** **on** 'A bear attacked a fisherman'	**NOM_DATvra**

=> Eastern Armenian equivalents of 'reach' and 'touch' belong to the same class; the equivalent of 'attack' is different

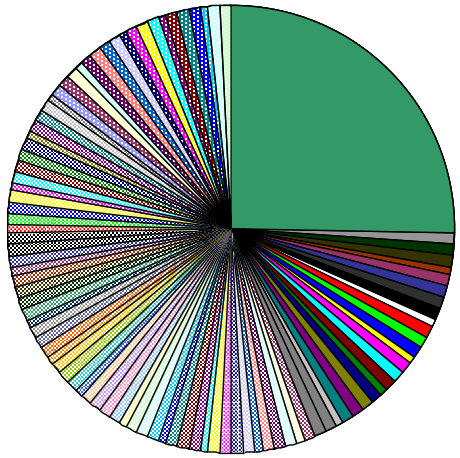
DistValPat

$$H(x) = -\sum_{i=1}^{k} p(x_i) \cdot \log(p(x_i))$$



Hypothetical Language 1: All verbs belong to the same class

$$H = 0$$

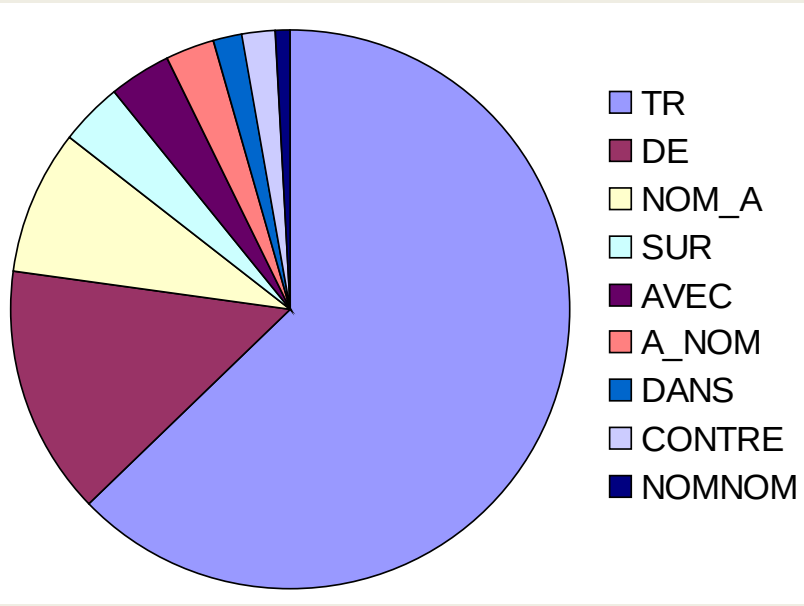


Hypothetical Language 2: 130 verb classes

$$H = \log(\frac{1}{130}) \approx 4{,}87$$

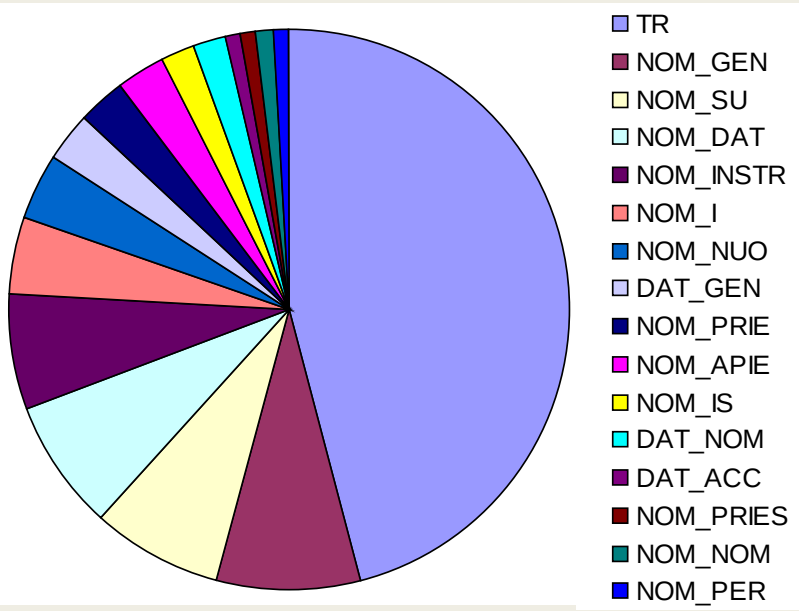
DistValPat

French



H ≈ 1.31

Lithuanian



H ≈ 2.02

	Armenian	Azerbaijani
take	TR	TR
see	TR	TR
influence	NOMvra	NOMDAT
encounter	TR	NOMCOM
enter	NOMNOM	NOMCOM
win	TR	NOMDAT
go_out	NOMABL	NOMABL
drive	TR	TR
bend	TR	TR
tell	NOMDAT	TR
hold	TR	TR
catch_up	NOMDAT	NOMDAT
milk	TR	TR
reach	NOMDAT	NOMDAT
touch	NOMDAT	NOMDAT
fight	NOMhet	NOMCOM
be_friends	NOMhet	NOMCOM
think	NOMmasin	NOMABL
…		
H (Entropy)	1.658	1.462

	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)
- MI (Armenian, Azerbaijani) = 1.658 + 1.462 − 2.196 = 0.924
- Higher MI values reflect higher similarity between valency class systems in the two languages
- MI was calculated using R package `infotheo` (Meyer 2014)

DistValPat

- 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

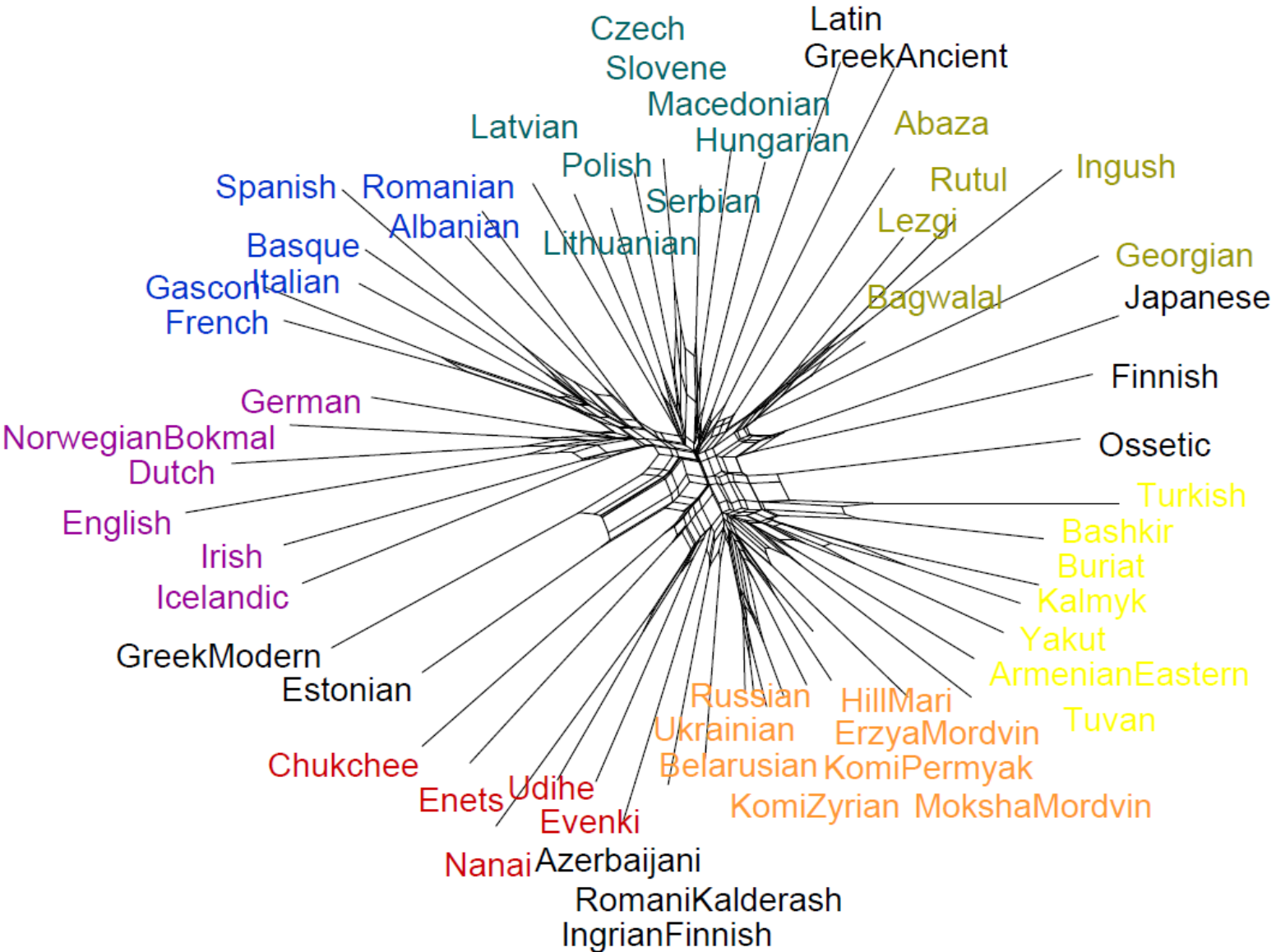
DistValPat

- DistValPal (Armenian, Azerbaijani) = 0.405

$$z\,(\text{DistValPat(Armenian, Azerbaijani)}) = \frac{0.405-0.499}{0.096} = -0.97$$

- The DistValPat distance between the two languages is almost one standard deviation below the mean => similar valency class systems

DistValPat: Neighbornet (select languages)



DistValPat: Neighbornet visualization

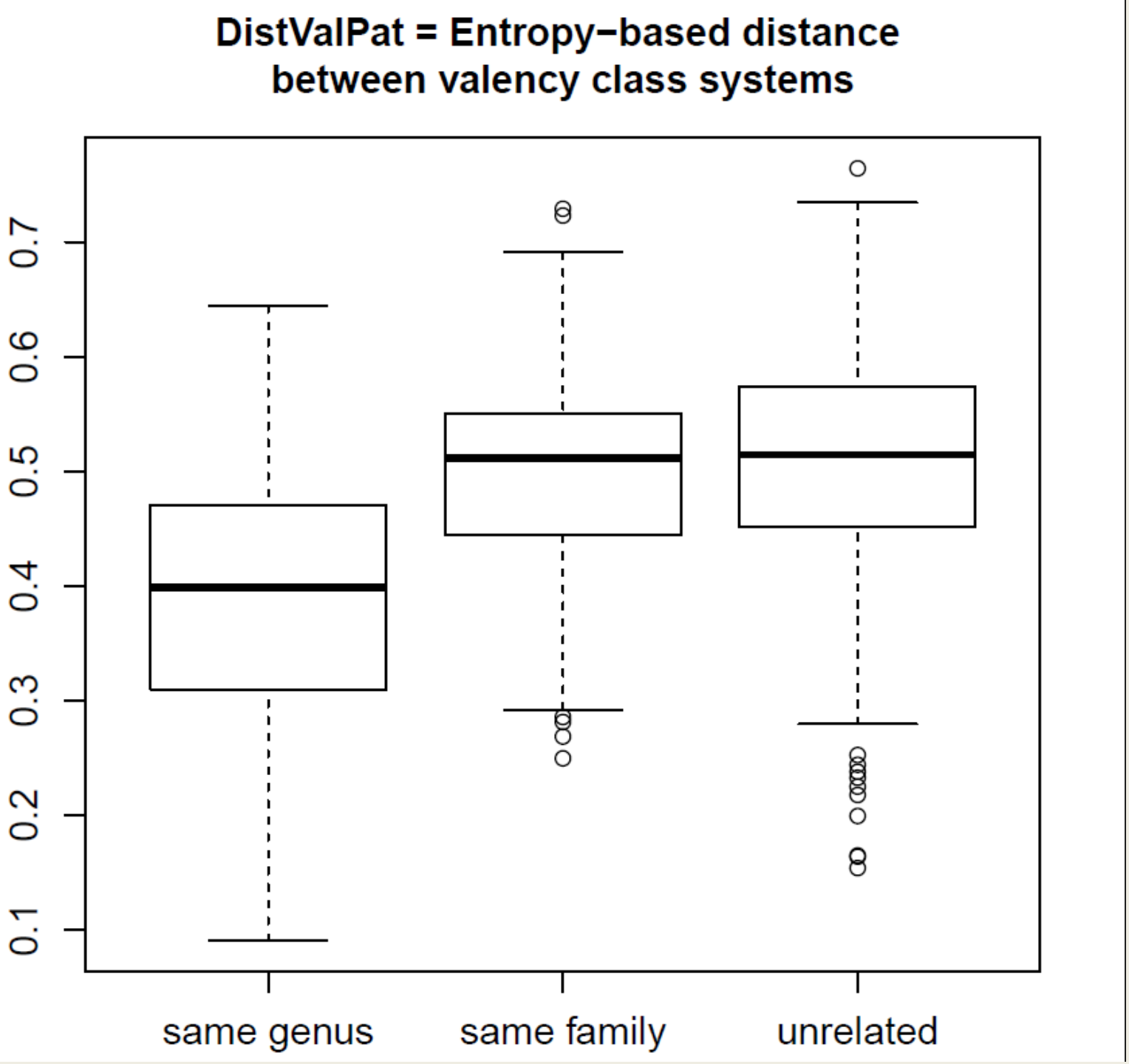
- Preliminary observations:
  - No one-dimensional structure visible
  - Many small areal clusters

- But this is an impressionistic observation
- And the genealogical dimension is not factored out

  => *see the next section*

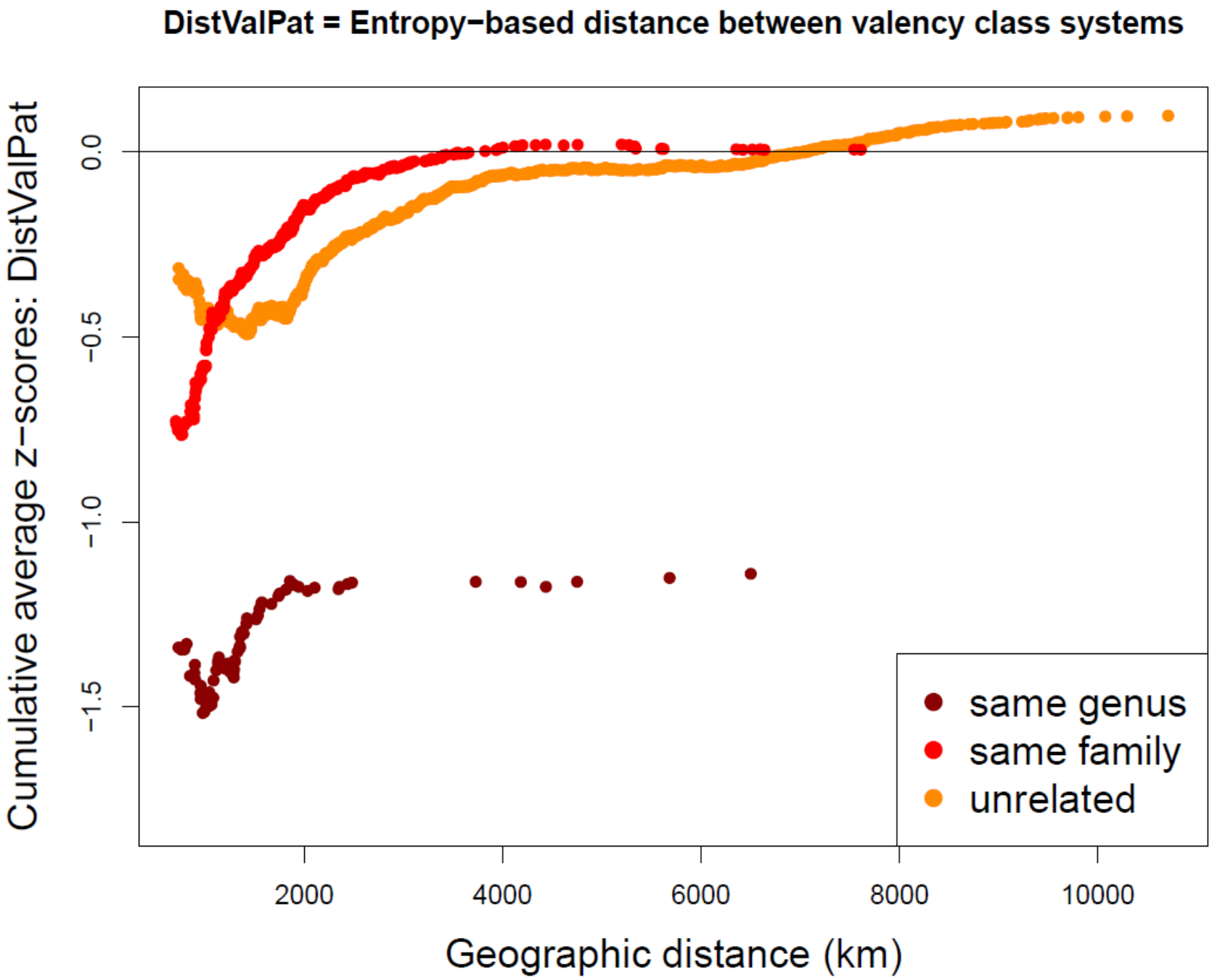
DistValPat

- Organization of low-level valency classes displays no family-level effects, only genus-level effects


DistValPat = Entropy−based distance
between valency class systems
0.1
0.2
0.3
0.4
0.5
0.6
0.7
same genus
same family
unrelated

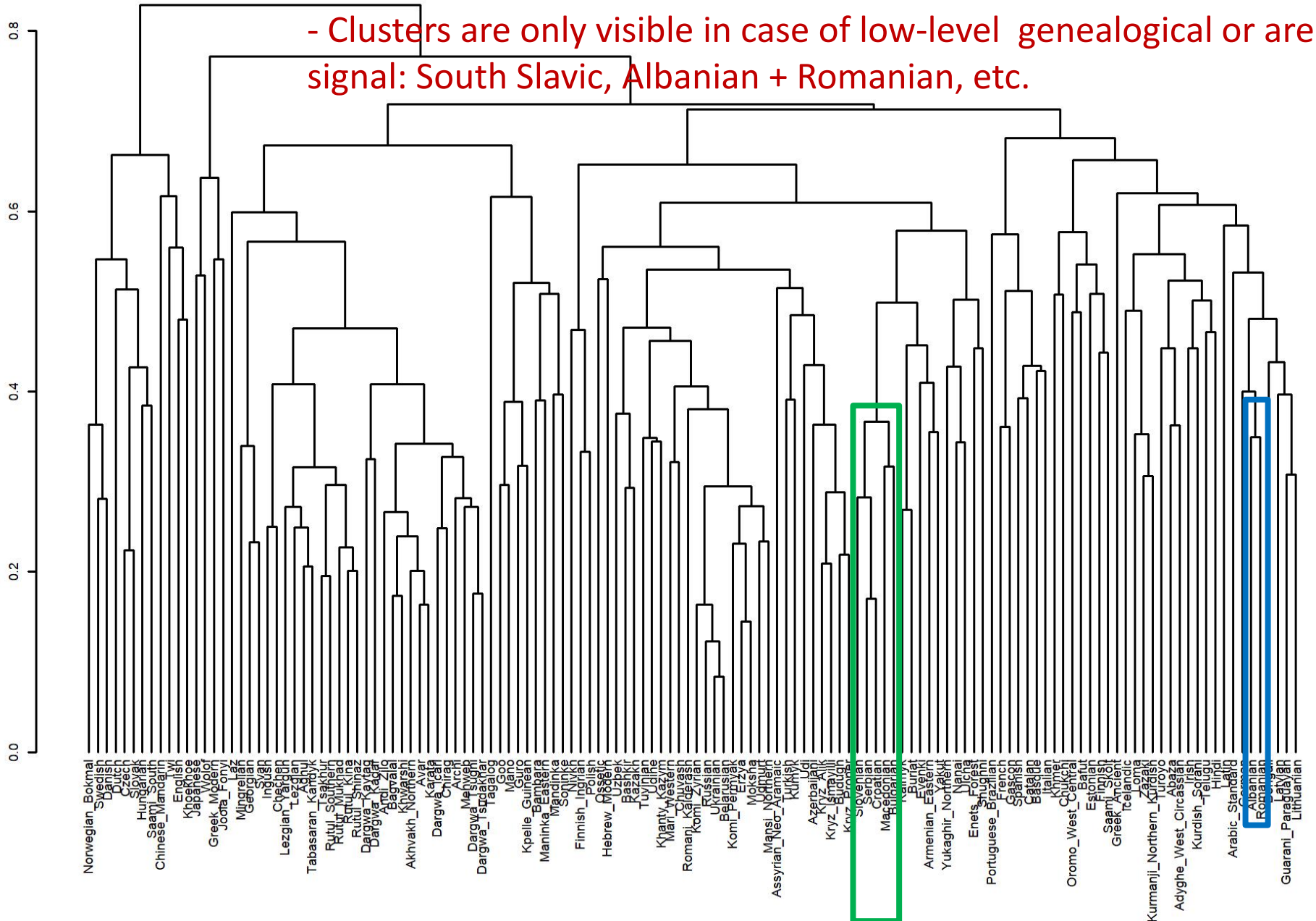
DistValPat

- DistValPat: geographical effects (next slide)
  - The curves for languages from same vs. different families show no consistent effect for distances > 1000 km
  - DistValPat shows the strongest areal signal for both genetically related and unrelated languages
  - Caucasus is an exception: many pairs of geographically proximate languages with huge DistValPat; this accounts for the anomaly on the left margin of the orange curve


DistValPat = Entropy−based distance between valency class systems
Cumulative average z−scores: DistValPat
0.0
−0.5
−1.0
−1.5
2000
4000
6000
8000
10000
Geographic distance (km)
same genus
same family
unrelated

Hierarchical clusters based on DistValPat

- Clusters are only visible in case of low-level genealogical or areal signal: South Slavic, Albanian + Romanian, etc.



Structure of the talk

- Setting the stage: typological study of valency
- The database: BivalTyp
- Cross-linguistic comparison
  - Valency patterns cognacy
  - Transitivity prominence
  - Lexical transitivity profiles
  - Lexical locus-based profiles
  - Lexical distributions into language-specific valency classes
- **Conclusions**

Conclusions

- Transitivity prominence is an areal phenomenon with subcontinental granularity
- Similarities in transitivity profiles
  - family-deep genealogical effects
  - weak large-scale areal effects

Conclusions

- Locus-based similarities
  - relatively stable genealogically
  - local areal convergences
- Similarities in the lexical organization of valency classes
  - no family-level genealogical effects
  - strong local areal convergences

Conclusions

- Plausible explanation
  - Valency patterns of individual verbs change relatively fast and are easily transferable in language contact
  - Languages are relatively stable in terms of those semantic features that are relevant for the assignment of the [+/-] transitivity values to individual verbs
  - Transitivity hierarchies of verb meanings can be family-specific

THANK YOU!

Selected references

- Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt & Alena Witzlack-Makarevich, 2014. Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment. Studies in language, 38 (3). Advances in research in semantic roles. 485-511.
- Bossong, Georg. 1998. Le marquage de l'expérient dans les langues d'Europe'. In: Feuillet (ed.). Actance et valence dans les langues de l'Europe. Berlin: Mouton de Gruyter. 259–94.
- Haspelmath, Martin. 1993. More on the typology of inchoative / causative verb alternations. In: Comrie, Bernard & Maria Polinsky (eds.) *Causatives and Transitivity.* Amsterdam: Benjamins. 87–120.
- Nedjalkov, V.P. 1969. Nekotorye verojatnostnye universalii v glagolnom slovoobrazovanii. In: F. Vardul' (ed.). Jazykovye universalii i lingvisticheskaja tipologija. Moscow: Nauka. 106-114.
- Nichols, Johanna. 2008. Why are stative-active languages rare in Eurasia? Typological perspective on split subject marking. In Mark Donohue and Søren Wichmann (eds). The Typology of Semantic Alignment Systems, 121-139. Oxford: Oxford University Press.
- Nichols, Johanna, David A. Peterson & Jonathan Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8: 149–211.
- Tsunoda, T. 1981. Split case-marking patterns in verb-types and tense / aspect / mood // Linguistics. Vol. 19. P. 389–438.
- Tsunoda, Tasaku. 1985. Remarks on transitivity. *Journal of Linguistics* 21. 385–396.