

Slavic Indefinite Pronouns from Multiple Wh-Words: A Corpus-Based Study of Grammaticalization, Semantics, and Distribution



Universität Regensburg

Dr. Roman Fisun
Institut für Slavistik
Fakultät Sprach-, Kultur- und Literaturwissenschaften

SLE
Osnabrück, 26.08.2026

Structure

1. Slavic indefinite series from double wh-constructions
2. Data and research questions
3. Inventory
4. Grammaticalization
5. Semantics vs. Distribution
6. Conclusion

Project background



Data and analyses from the completed DFG-funded project

“Compound Indefinite Pronouns in Slavonic Languages: A Contribution to a Second-Generation Semantic Map of Indefiniteness” (HA 2659/10-1)

Languages: Russian, Ukrainian, Polish, Czech and Croatian

PI: Prof. Dr. Björn Hansen

Scope: weakly grammaticalized and non-codified compound indefinite pronouns in Slavic

Most Croatian data annotated by Björn Hansen, most Czech data by Martina Rybová.

Project materials and datasets:

<https://fisun.org/files/diss/datasets/>



The phenomenon: Indefiniteness marker derived from an interrogative stem

- Typical pattern for indefinite pronouns in Slavic languages:
series consisting of an indefiniteness marker and a variable pre- or postposed interrogative stem (*k-root)
 - One source of indefinite markers: the Proto-Slavic pronominal interrogative *k-root
 - Paradoxical structure: [*k-root as marker] + [*k-root as interrogative stem]
- HR **gdje**-tko_{where-who} 'someone', **gdje**-što_{where-what} 'something', **gdje**-koji_{where-how} 'some'
- UA **de**-hto_{where-who} 'someone', **de**-ščo_{where-what} 'something', **de**-jak_{where-how} 'somehow'
- One former wh-element becomes fixed as an indefiniteness marker, the other remains paradigmatically variable

Research questions

- (i) Which Slavic indefinite-pronoun series contain **k*-root-derived indefiniteness markers?
- (ii) How far are the corresponding markers grammaticalized?
- (iii) Do the series converge on a shared semantic profile?
- (iv) Do semantic functions show systematic preferences across licensing contexts?

Data

5 Slavic languages: Russian, Ukrainian, Polish, Czech, Croatian

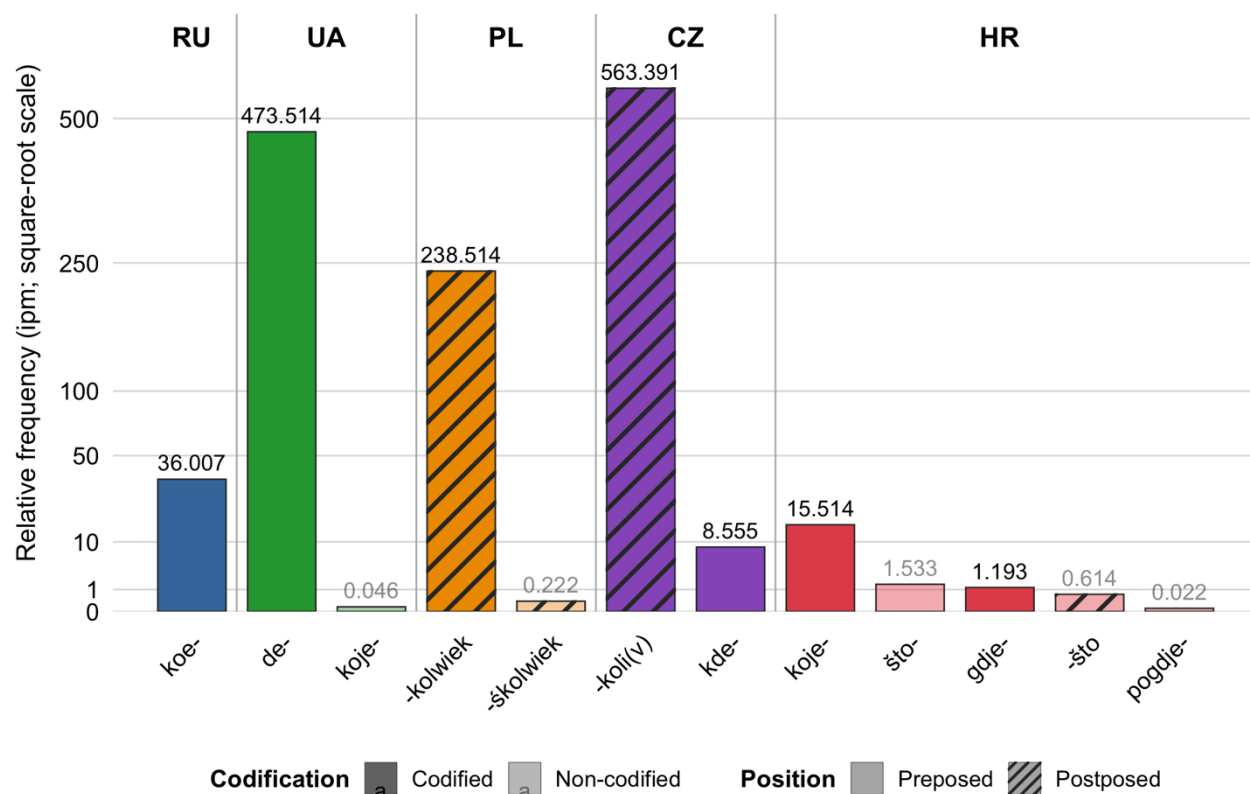
Large web corpora: TenTen family and hrWaC 2.2, available on Sketch Engine

Why web corpora? The relevant forms are often rare, variable, or absent from normative descriptions

Language	Corpus	Tokens
RU	ruTenTen11	18.3 bn
UA	ukTenTen20	3.3 bn
PL	plTenTen12	9.4 bn
CZ	csTenTen17	12.6 bn
HR	hrWaC 2.2	1.4 bn

Inventory: 12 series with **k*-root-derived indefiniteness markers

Relative Frequency of **k*-Root Indefinite Series



Identification in Sketch Engine

- Word List for single-word and hyphenated forms
- collocational searches for potentially separate elements

12 series: 1 RU · 2 UA · 2 PL · 2 CZ · 5 HR

All markers in this set form one orthographic word with the interrogative stem

Two structural groups

Group 1 PL *-(ś)kolwiek*, CZ *-koli(v)*

- postposed
- additional material beyond the wh-element
- etymologically less transparent

Group 2 RU *koe-*, UA *de-*, *koje-*, CZ *kde-*, HR *gdje-*, *koje-*, *pogdje-*, *što-*, **-što**

- predominantly preposed

Measuring Grammaticalization

Lehmann (2015 [1982]): six parameters of grammaticalization

Paradigmatic axis

Integrity ↓

Paradigmaticity ↑

Paradigmatic variability ↓

Syntagmatic axis

Structural scope ↓

Bondedness ↑

Syntagmatic variability ↓

Operationalization: 15 corpus-based subcriteria (Fisun & Hansen 2025; Fisun 2026)

Evidence: quantitative + qualitative corpus analysis

Scoring: each parameter scored 0–1; overall score = mean of the six parameter scores

Descriptive ranges: weak $\leq .33$ · medium $.34\text{--}.66$ · strong $\geq .67$

E. g., Paradigmaticity: 1. Substitutability with a codified indefinite pronoun (in at least some contexts or in translation). 2. Range across ontological categories (≥ 3 different interrogative stems)

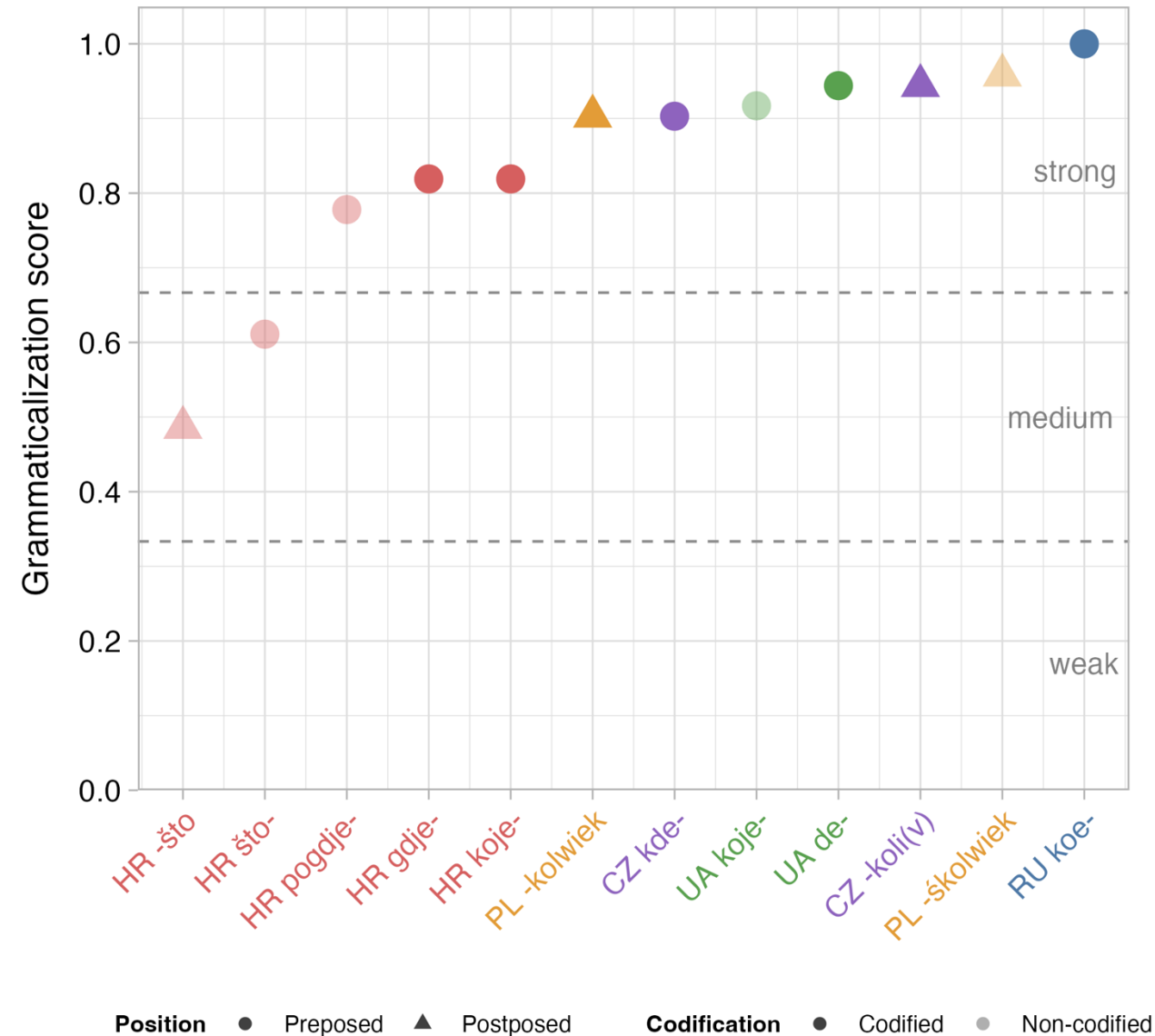
Grammaticalization score

Main result

- Most series show a high degree of grammaticalization
- 10 of 12 series fall into the strong range
- Only 2 series remain in the medium range — both Croatian
- The scores do not reproduce the source-based structural grouping

Source type ≠ grammaticalization degree

Grammaticalization Scores of *k-Root Indefinite Series



Revisited semantic maps

Two dimensions kept separate

1. semantic function of the indefinite form
2. licensing context in which the function is realized

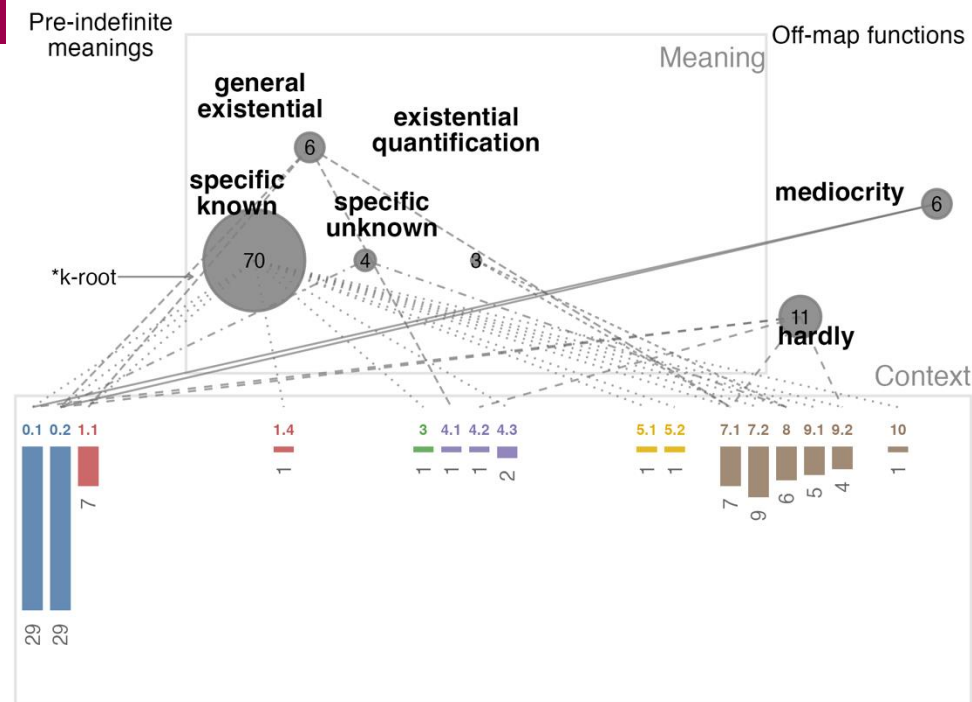
Semantic domain

- **Core semantic functions:** *specific known* · *specific unknown* · *general existential* · *existential quantification* · *free choice* · *referential conditional* · *interrogative*
- **Off-map functions** (Aguilar-Guevara et al. 2011): non-pronominal meanings, often evaluative or lexicalized

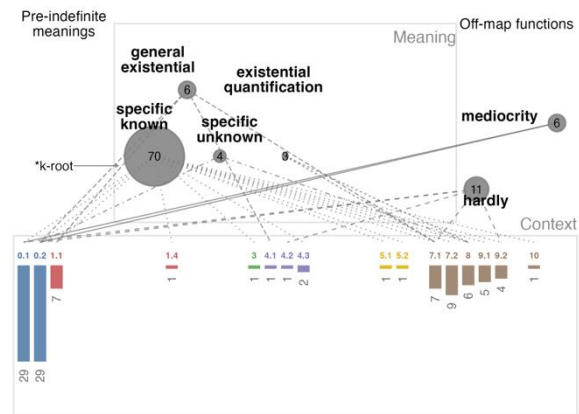
Distributional domain

- **Licensing contexts:** 9 broad groups for the presentation; >27 subtypes in the full annotation scheme; framework: suspended assertion / non-veridicality (Padučeva 2015, 2018; Giannakidou 1998)

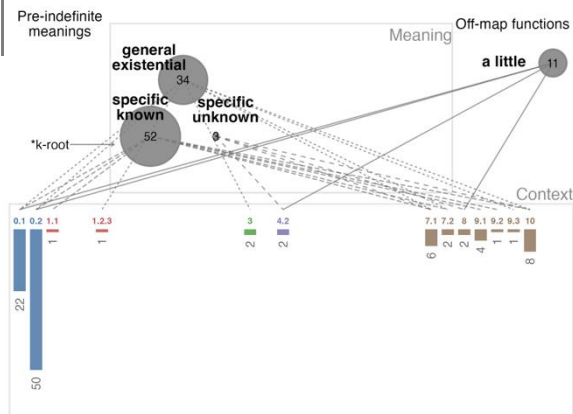
Data: target sample of 100 corpus occurrences per series, manually annotated. HR *pogdje*- uses all 44 available examples, normalized to 100 for comparison



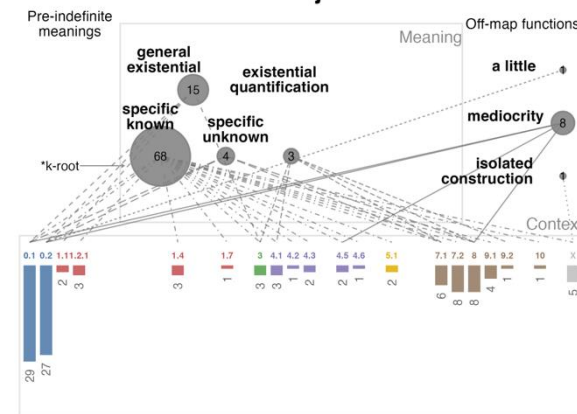
RU koe n=100



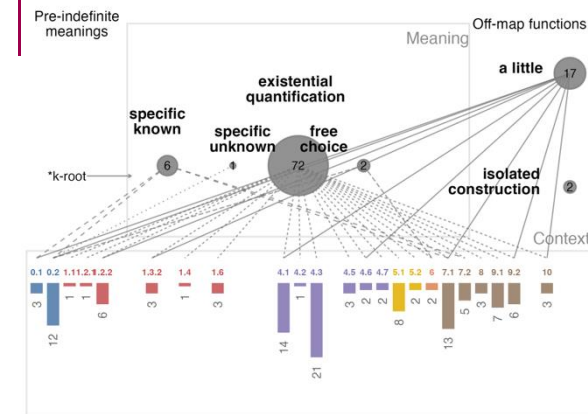
UA de n=100



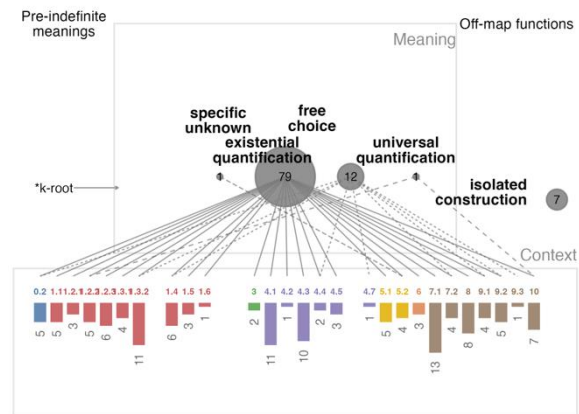
UA koje n=100



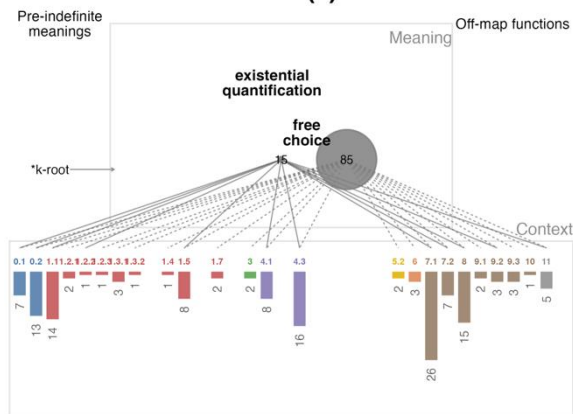
PL školwiek n=100



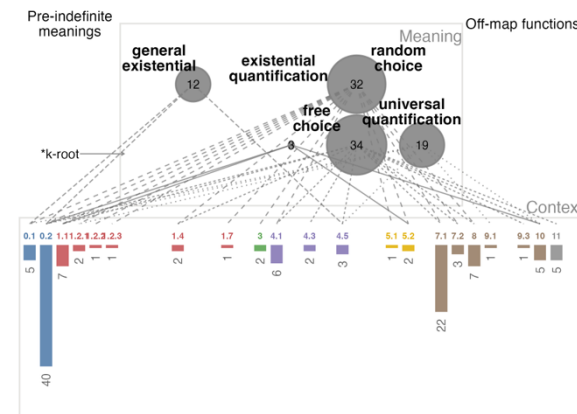
PL kolwiek n=100



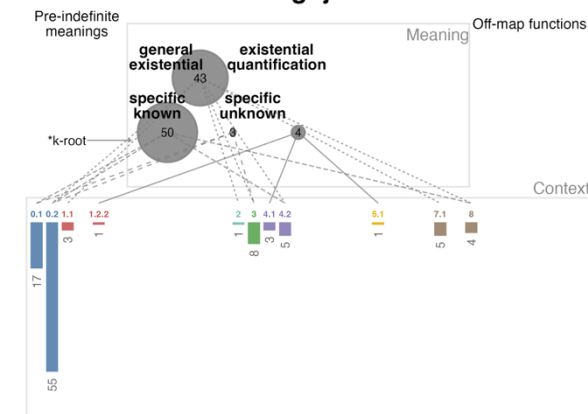
CZ koli(v) n=100



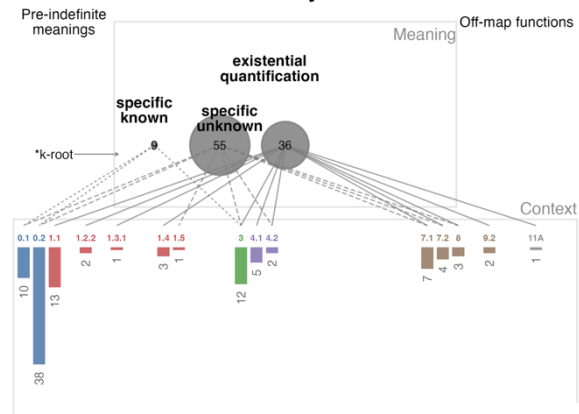
CZ kde n=100



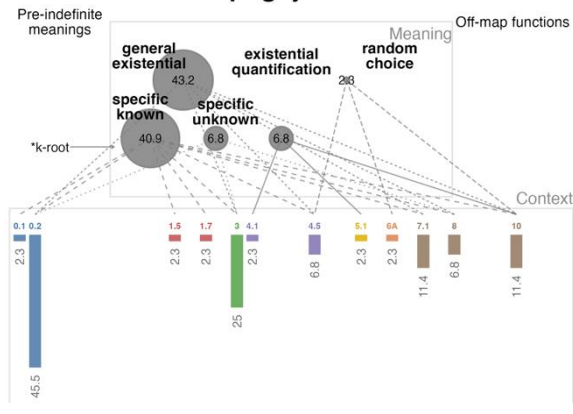
HR gdje n=100



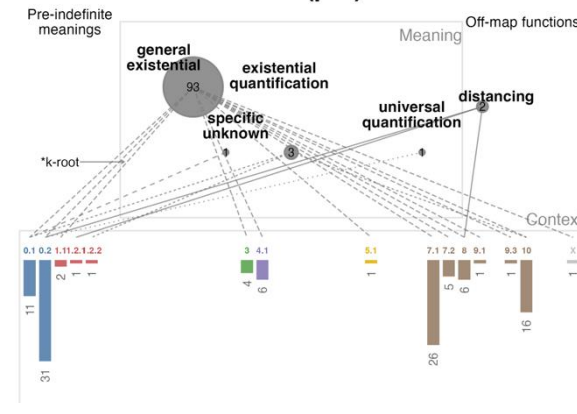
HR koje n=100



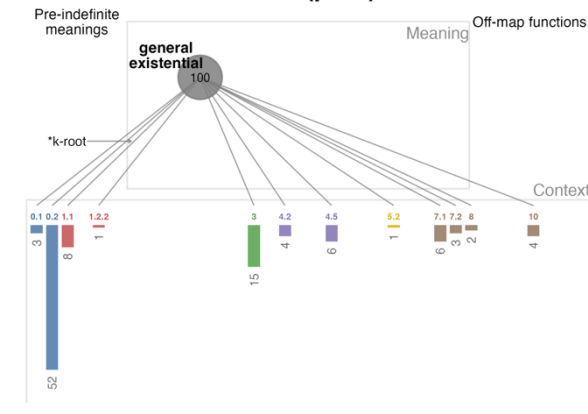
HR pogdje n=44->100



HR što (pre) n=100



HR što (post) n=100



Assertive Negation Existential set Iterative Conditional Questions Disjunctions Irrealis Comparison

Examples

specific known:

(1) *Prežde my dolžny **koe-čto** obsudit'... ja dolžna byt' uverena.* (ruTenTen11)

'First, we have something to discuss ... I have to be sure.'

off-map mediocrity:

(2) *Fred **koe-kak** otrjachnulsja i dvinulsja dal'se.* (ruTenTen11)

'Fred dusted himself off rather poorly and moved on.'

general existential:

(3) ***Dejaki** ljudi, chvori hrypom, možut' vyprovovuvaty til'ky **dejaki** z cych symptomiv.*

(ukTenTen20)

'Some people with influenza may experience only some of these symptoms.'

random choice:

(4) *Dialogy zde si může vymyslet **kdejaký** škrabácký moul.* (csTenTen17)

'The dialogue here can be made up by any random scribbling fool.'

Weighted semantic map (Cysouw 2007)

Series-based map – top

node size = number of series with the function

edge width = number of series with the connection

Frequency-based map – bottom

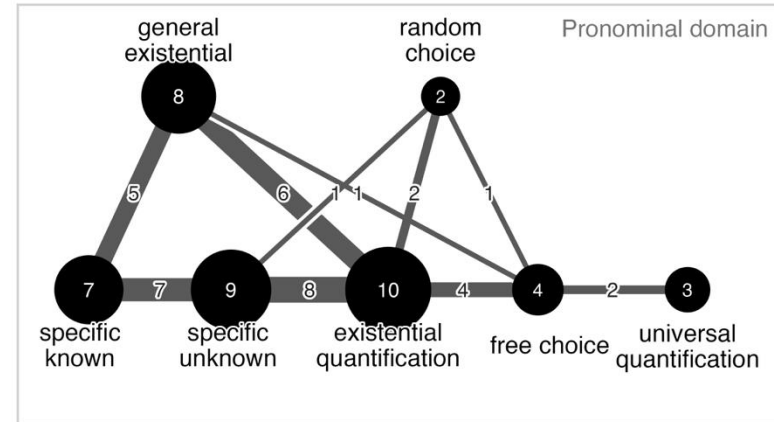
node size = total token frequency across all 12 series

edge width = frequency-weighted connection strength

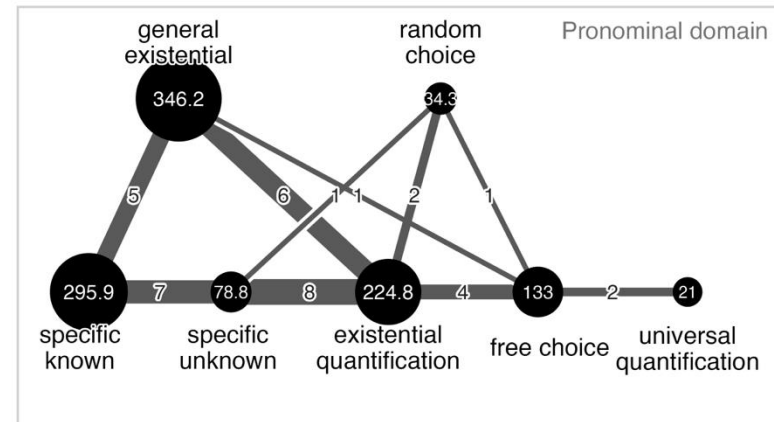
Main pattern

- *existential quantification* is most widespread across series
- *general existential* and *specific known* dominate token frequency
- *free choice* is frequent, but concentrated in fewer series

Type II – series-based



Type II – frequency-based

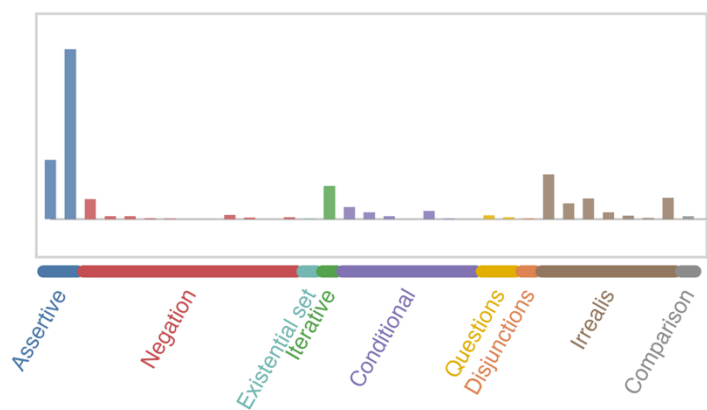
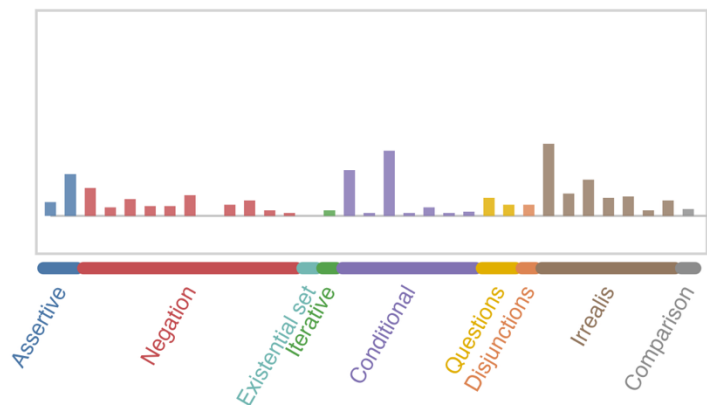


Cluster analysis of semantic and distributional profiles

- Goal: identify groups of series with similar semantic or distributional profiles
- Method: hierarchical clustering using Euclidean distance + Ward.D2
- Silhouette identifies the best-supported number of clusters
 - Distribution: $k = 2$, silhouette = .381
 - Semantics: $k = 5$, silhouette = .570
- Results: distribution forms two broad groups, while semantic profiles are more internally differentiated
- The two distributional clusters exactly match the source-based grouping: ARI = 1.00
- For semantics, a forced two-cluster solution shows only partial agreement with the source-based grouping: ARI = .40

Distribution

$k = 2$ · silhouette = .381 · ARI = 1.00



CZ -koli(v)

PL -kolwiek

PL -śkolwiek

CZ kde-

HR gdje-

HR koje-

HR pogdje-

HR što-

HR -što

RU koe-

UA de-

UA koje-

CZ kde-

CZ -koli(v)

HR gdje-

HR pogdje-

RU koe-

UA de-

UA koje-

HR koje-

HR što-

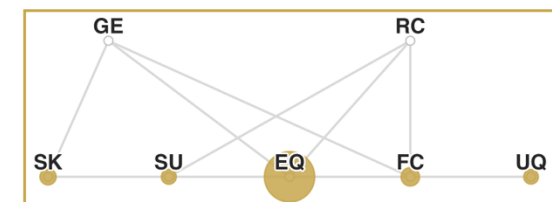
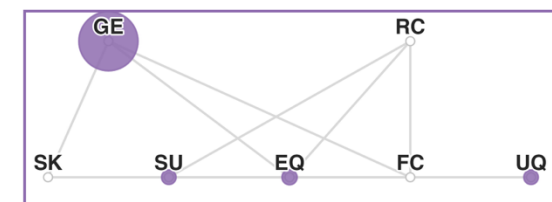
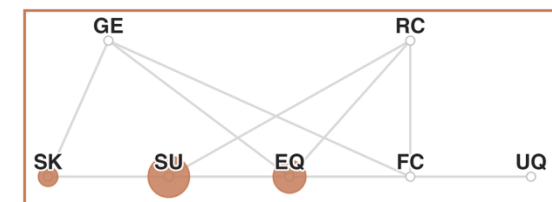
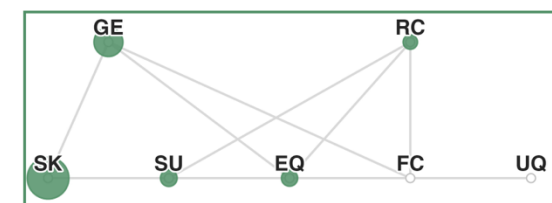
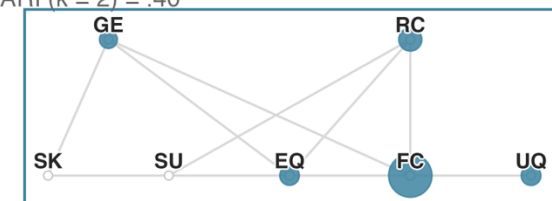
HR -što

PL -kolwiek

PL -śkolwiek

Semantics

$k = 5$ · silhouette = .570 · ARI ($k = 2$) = .40



Conclusion

- 12 **k-root* series across five Slavic languages, two main source groups
- Grammaticalization: generally high, does not follow the source grouping
- Distribution: two clusters reproduce the source groups exactly
- Semantics: best-supported solution = five clusters; much stronger differentiation
- Semantic core: specific reference, general existential, and existential quantification
- Shared source strongly predicts distributional similarity, but not uniform semantic or grammaticalization outcomes

References

- Aguilar-Guevara, A., M. Aloni, A. Port, R. Šimík, M. de Vos & H. Zeijlstra (2011) Semantics and pragmatics of indefinites: Methodology for a synchronic and diachronic corpus study. In: S. Dipper & H. Zinsmeister (eds.) *Beyond Semantics: Corpus-Based Investigations of Pragmatic and Discourse Phenomena. Proceedings of the DGfS Workshop Göttingen, February 23–25, 2011*, 1–16. Bochum
- Apresjan, V. & Kopotев, M. (2022). Avtonomnye distributivnye konstrukcii s voprositel'no-odnositel'nymi mestoimenijami v russkom jazyke. *Voprosy jazykoznanija* 4, 115–142.
- Citko, B. (2010). On the distribution of -kolwiek 'ever' in Polish free relatives. *Journal of Slavic Linguistics* 18(2), 221–258.
- Cysouw, M. (2007). Building semantic maps: The case of person marking. In M. Miestamo & B. Wälchli (eds.), *New Challenges in Typology*, 225–248. Berlin: Mouton de Gruyter.
- Fisun, R. (2026). Zusammengesetzte Indefinitpronomen in den slavischen Sprachen. Regensburg. [PDF](#)
- Fisun, R., & Hansen, B. (2025). A corpus-based study on the inventory and grammaticalisation of compound indefinite pronouns in Croatian. *Die Welt der Slaven*, 69(2), 293–330. [PDF](#)
- Giannakidou, A. (1998) *Polarity Sensitivity as (Non)veridical Dependency*. Amsterdam
- Haspelmath, M. (1997). *Indefinite Pronouns*. Oxford: Oxford University Press.
- Lehmann, C. (2015). *Thoughts on Grammaticalization*. 3rd ed. Berlin: Language Science Press.
- Padučeva, E. V. (2015). Snjataja utverditel'nost' i neveridiktal'nost' (na primere russkich mestoimenij otricateľ'noj poljarnosti). *Russian Linguistics* 39(2), 129–162.
- Padučeva, E. V. (2017). Referencialnyj status imennoj gruppy. *Russkaja korpusnaja grammatika*.
- Padučeva, E. V. (2018) Russkie mestoimenija svobodnogo vybora. *Russian Linguistics* 42(3), 291–319

Thank you for your attention!

Contact: roman@fisun.org

Materials and raw data:
<https://fisun.org/files/conf/SLE26/>

