Measuring the language distance between the lects with high degree of inner variation (on the material of South Slavic lects)

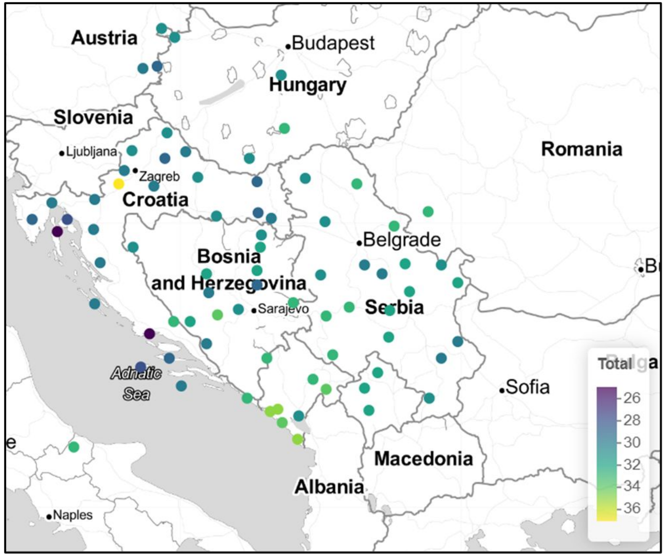
Ilia Afanasev (HSE University/MTS AI),
Viacheslav V. Kozak (Institute for Linguistic Studies RAS)

Talk structure

- Introduction: recap & material characterisation
- Problem stated
- Methodology outline
- Experiments & analysis
- Conclusion & further directions

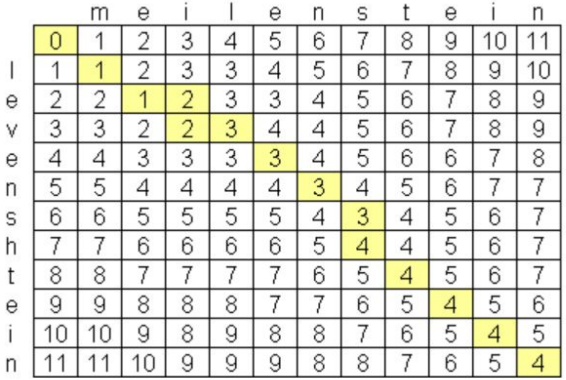
Short recap

- Corpus-based study of language distance: investigating, whether it is possible to build a preliminary (genetic) classification of languages, relying on raw (completely unprocessed) corpus data
- Documenting South Slavic lects
- Testing phylogenetic methods: borrowing methods from computational biology and test, whether they suit a particular research goal



Broader task

- Assemble a 40-item (Holman et al., 2008) Swadesh list for a set of South Slavic lects
- Build a preliminary consensus tree with Levenshtein distance normalised divided (LDND; Holman et al., 2008) and weighted Jaro-Winkler distance normalised divided (WJWDND; Gueddah et al., 2015)
- Use one more lect as an outgroup (cf. Kassian et al., 2021) to build a more precise internal classification



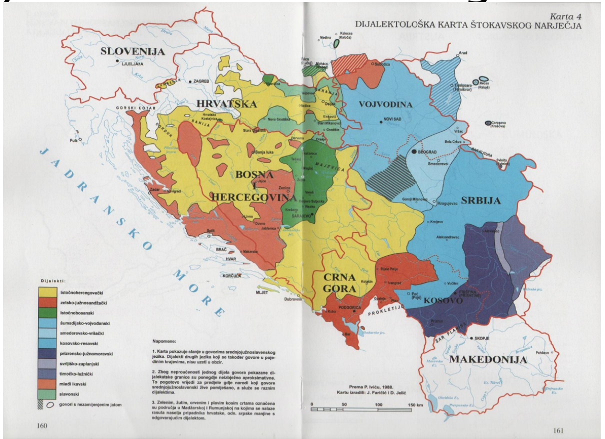
	m	e	i	l	e	n	s	t	e	i	n
0	1	2	3	4	5	6	7	8	9	10	11
l 1	1	2	3	3	4	5	6	7	8	9	10
e 2	2	1	2	3	3	4	5	6	7	8	9
v 3	3	2	2	3	4	4	5	6	7	8	9
e 4	4	3	3	3	3	4	5	6	6	7	8
n 5	5	4	4	4	4	3	4	5	6	7	7
s 6	6	5	5	5	5	4	3	4	5	6	7
h 7	7	6	6	6	6	5	4	4	5	6	7
t 8	8	7	7	7	7	6	5	4	5	6	7
e 9	9	8	8	8	7	7	6	5	4	5	6
i 10	10	9	8	9	8	8	7	6	5	4	5
n 11	11	10	9	9	9	8	8	7	6	5	4

Terminological clarification

- Lect is any given variety of the language, such as:
    - idiolect
    - doculect
    - dialect
    - sociolect
    - standard
- The term is introduced to reduce the possible synchronous hierarchy discussions (“language - dialect” problem, cf. Koryakov, 2017; Fedotova, 2022)

South Slavic lects - material

1. “Hvar” — Southern Čakavian dialect of Hvar (Benčić 2014).
2. “Kuči” — Zeta-Lovćen Štokavian dialect of Kuči, Eastern Montenegro (Петровић, Ћелић, Капустина 2013).
3. “North Metohija” — Kosovo-Resava Štokavian dialect of the North Metohija region (Букумирић 2012).
4. “Lužnica” — Prizren–Timok Štokavian (or Torlak) dialect of Lužnica region (Ћирић 2018).



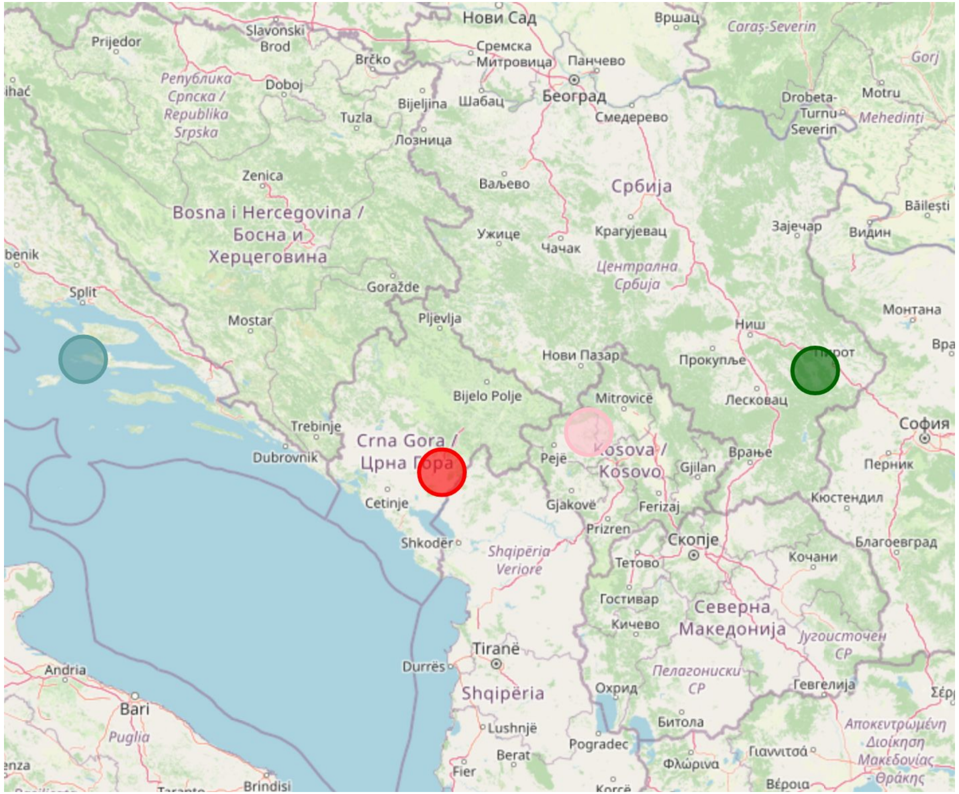
Napomene:

1. Karta pokazuje stanje u govorima srednjojužnoslavenskog jezika. Dijalekti drugih jezika koji se također govore u pojedinim krajevima, nisu uzeti u obzir.

2. Zbog neproučenosti jednog dijela govora pokazane dijalekatske granice su ponegdje neizbježno aproksimativne. To pogotovo vrijedi za predjele gdje narodi koji govore srednjojužnoslavenski žive pomiješano, a služe se raznim dijalektima.

3. Zelenim, žutim, crvenim i plavim kosim crtama označena su područja u Mađarskoj i Rumunjskoj na kojima se nalaze rasuta naselja pripadnika hrvatske, odn. srpske manjine s odgovarajućim dijalektom.

South Slavic lects - map



● Kuči

● North Metohija

● Lužnica

● Hvar

Complications

- Lects possess a high degree of lexical and/or phonetic variation, most notably:
  - Words of historically different roots that represent the single concept within the single lect: *nidra, parsi, sisa* ‘breast’ (Hvar)
  - Words of historically same root but the different phonetic form that represent the single concept within the single lect: *kos, koska, koʧina, koʃʧina* ‘bone’ (Lužnica)
- With the existing material, it is hard to search for a diagnostic contexts and apply rigorous enough criteria (Kassian et al., 2010; Afanasev, 2023)
- This heavily complicates the use of string similarity measures

36. TREE

- Kuči: *drijevo, drvo*
- Lužnica: *drvo*
- North Metohija: *darva, drivo*
- Hvar: *drvo*

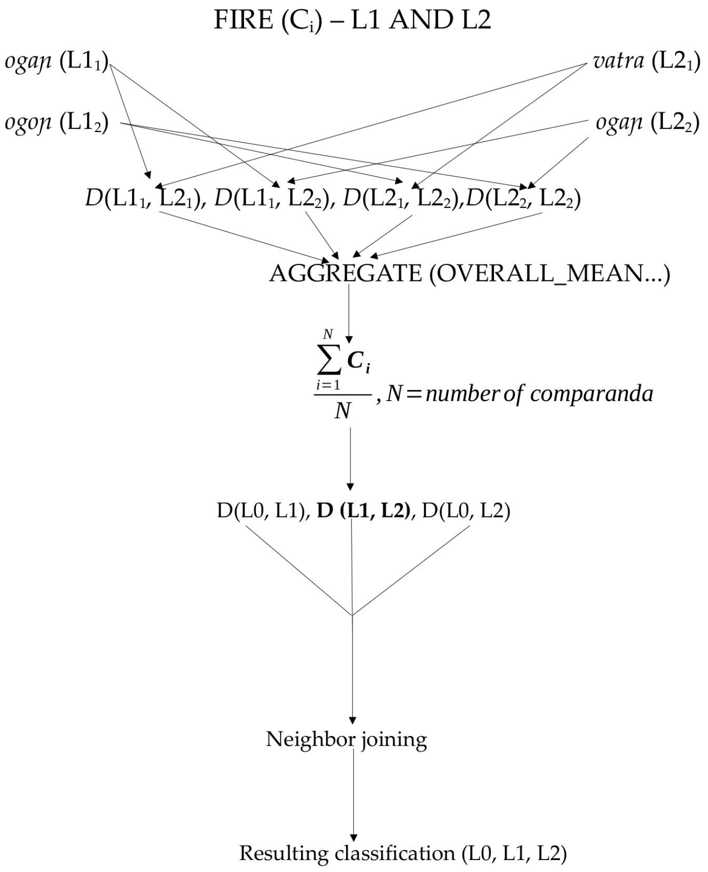
21. MOUNTAIN

- Kuči: *planina*
- Lužnica: *gora, planina*
- North Metohija: *planina*
- Hvar: *muntaɲa*

Possible solutions

- Use different combinations of minimal, mean and maximal values:
    - OVERALL_MEAN: Scoring mean among all the distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_m$ of concept C between lects A and B
    - MEAN_MEAN: Scoring mean among means of the distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_n$ of concept C between lects A and B
    - MEAN_MIN: Scoring mean among minimal distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_n$ of concept C between lects A and B
    - MEAN_MAX: Scoring mean among maximal distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_n$ of concept C between lects A and B
    - MIN_MEAN: Picking minimal value among means of the distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_m$ of concept C between lects A and B
    - MAX_MEAN: Picking maximal value among means of the distances (a_i, b_j) for all realisations $a_1…a_n$ and $b_1…b_m$ of concept C between lects A and B
- Use threshold for automatic non-cognate elimination, and pick minimal value afterwards

Workflow



FIRE (C_i) – L1 AND L2

ogaɲ ($L1_1$)

vatra ($L2_1$)

ogoɲ ($L1_2$)

ogaɲ ($L2_2$)

$D(L1_1, L2_1)$, $D(L1_1, L2_2)$, $D(L2_1, L2_2)$,$D(L2_2, L2_2)$

AGGREGATE (OVERALL_MEAN...)

$$\frac{\sum_{i=1}^{N} \boldsymbol{C_i}}{N}, N = number\, of\ comparanda$$

D(L0, L1), **D (L1, L2)**, D(L0, L2)

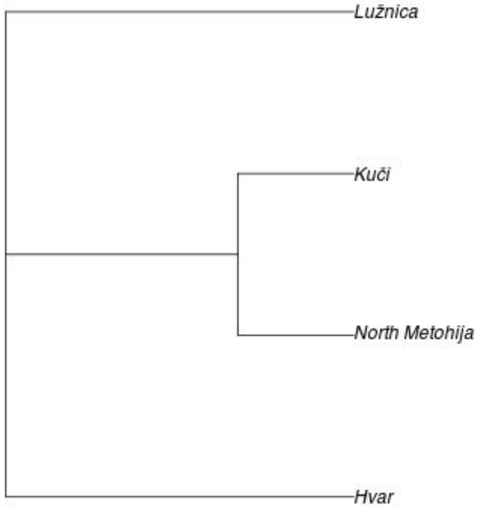
Neighbor joining

Resulting classification (L0, L1, L2)

Experiment flow

- Manually construct the gilded tree
- Use black-box approach (Afanasev, forthc.) as a baseline and to determine complexity of the task
- Score distances by each of the possible aggregated string similarity measures
- Score distances with thresholds of 0.1, 0.33, 0.5, 0.66 and 0.9 (from the most restrictive to the least restrictive)
- Evaluate results with mutual clustering information (MCI; Smith, 2020)
- Conduct a linguistic analysis of the pairs, rejected by given threshold

Gilded tree



Black-box method

- Includes elimination of transparency for both a researcher and explanatory methods (Munn and Pitman, 2022) via
  - Ciphering
  - BPE (byte-pair encoding) tokenisation
  - Vectorisation
  - Classification of concepts by lects with Random Forest Classifier (Ho, 1995)
  - Measuring the distance by loss in mean square error between initial classification and classification after random swap of *M* concepts ($a_i, b_i > b_i, a_i$), with the possibility of imitating borrowing ($a_i, b_i > a_i, a_i$)
- Evaluated through mean MCI within *N* runs
- The least required (we use 1000) number of runs is calculated by formula: $1 - (S - M/S)^N > 0.999$, where *S* is number of concept in given comparison, *M* is number of swapped concepts, and *N* is a number of runs

Black-box method results

Experiment	Number of swaps	Borrowing	Mean MCI (1000 runs)
1	3	0	0.339
2	3	1	0.359
3	14	0	0.475
4	14	1	0.463

Black-box method analysis

- The automatic methods are applicable to the task
- The complexity is higher, than for classifying three East Slavic lects of slightly more shallow relationship (average probability of acquiring correct tree ~= 0.6), and equals to classifying Taa (average probability of acquiring correct tree ~= 0.4; Afanasev, forthc.)

String similarity measures: naive approaches

Experiment	MCI
OVERALL_MEAN	1
MEAN_MEAN	1
MEAN_MIN	1
MEAN_MAX	1
MIN_MEAN	1
MAX_MEAN	1

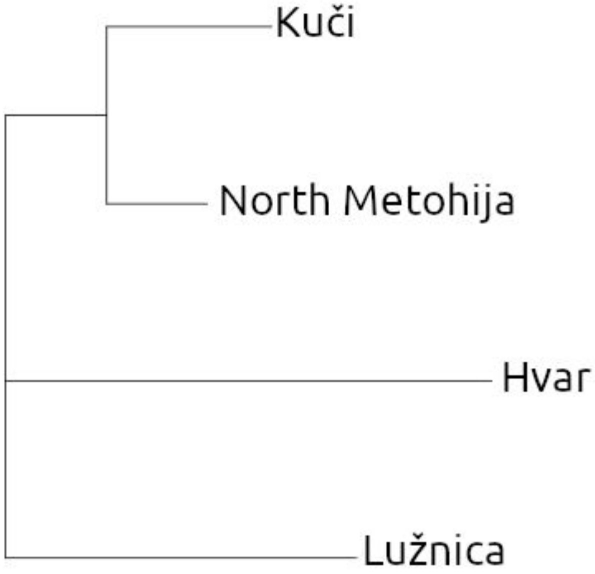
String similarity measures: introducing threshold

Threshold	MCI
0.1	0
0.33	0
0.5	1
0.66	1
0.9	1

String similarity measures: detected (non-)cognates

Concept	Word pair	Lect pair	Threshold	Value
EAR	*uo/uvo*	Kuči/Lužnica	0.1	0.34
HUMAN	*ʧek/ʧovek*	Kuči/North Metohija	0.33	0.4
LIVER	*ʤigeriʦa/utrobiʦa*	Kuči/Lužnica	0.5	0.625
MOUNTAIN	*gora/muntaɲa*	Lužnica/Hvar	0.66	0.857
FIRE	*vatra/ogaɲ*	North Metohija/Hvar	0.9	1

Average tree (threshold = 0.5 & threshold = 0.66)



Preliminary results

- There is no huge difference between naive approaches and implementing threshold in terms of scores
- However, linguistic interpretability of results is significantly higher, when threshold is implemented
- When threshold is too low, a lot of cognates do not pass, which leads to incorrect results
- When threshold is too high, a lot of non-cognates pass through, which creates noise in data
- It seems that optimal threshold is approximately in [0.4; 0.6] interval

Future directions

- Clear dataset further according to guidelines in Kassian et al. (2010)
- Cross-verify with other data
- Cross-verify with WJWDND such cases as *ʤigeriʦa/utrobiʦa* ‘liver’
- Use a more probabilistic approach
- Use an outgroup method and determine, whether it yields more precise classification
- Test a similar approach against corpus data
- Collect 110-item wordlists for given lects

Thank you!

Fond de carte de l'Europe du Sud-Est, dressé initialement sur papier en 1987 par S. I. Cepleanu-Galtier étudiant à l'Inst. de Géographie de l'Université Paris IV-Sorbonne, et scanné en 2012 pour être téléchargé sur internet, par le même auteur devenu contributeur. Usage initial : S. I Cepleanu, *«Atlas géomorphologique des pays danubiens et balkaniques, mémoire de maîtrise de géographie»*.

References

Afanasev, Ilia. Forthcoming. Cipher, transform, get lost: an anti-transparent system for East Slavic lects distance measurement.
Afanasev, Ilia. 2023. Multi-lect automatic detection of Swadesh list items from raw corpus data in East Slavic languages. Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 76–86
Benčić, Radoslav. Rječnik govora grada Hvara : forske rici i šporije / Radoslav Benčić. - 2. prošireno i poboljšano izd.. - Hvar : Muzej hvarske baštine, 2014. - 527 str. : ilustr. ; 25 cm
Fedotova, Idaliya. 2022. Диалектное членение хантыйского языка по данным базисной лексики. Ural-Altaic Studies, 47, 117-166.
Gueddah, Hicham, Abdellah Yousfi, Mostafa Belkasmi. 2015. The filtered combination of the weighted edit distance and the Jaro-Winkler distance to improve spellchecking Arabic texts. 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA): 1-6.
Ho, Tin Kam. 1995. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition: 278–282.
Holman, Eric, Søren Wichmann, Cecil Brown, Viveka Velupillai, André Müller, Dik Bakker. 2008. Explorations in automated language classification. Folia Linguistica 42:331–354.
Kassian, Alexei, George Starostin, Anna Dybo, Vasiliy Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. Journal of Language Relationship, 16: 46–89.
Kassian, Alexei S., Zhivlov, Mikhail, Starostin, George, Trofimov, Artem A., Kocharov, Petr A., Kuritsyna, Anna and Saenko, Mikhail N.. "Rapid radiation of the inner Indo-European languages: an advanced approach to Indo-European lexicostatistics" Linguistics, vol. 59, no. 4, 2021, pp. 949-979. https://doi.org/10.1515/ling-2020-0060
Koryakov, Yu. B. 2017. Language vs dialect: a lexicostatistic approach, Voprosy Jazykoznanija , 6, 79—101.
Munn, Michael, David Pitman. 2022. Explainable AI for Practitioners. O'Reilly Media, Inc.
Smith MR (2020). "Information theoretic Generalized Robinson-Foulds metrics for comparing phylogenetic trees." *Bioinformatics*, **36**(20), 5007–5013. doi: 10.1093/bioinformatics/btaa614.
Букумирић, М.. (2012). Речник говора северне Метохије. Београд : Институт за српски језик САНУ.
Ћирић, Љ.. (2018). Речник говора Лужнице. in Српски дијалектолошки зборник. Београд : Институт за српски језик САНУ., 65(2).
Петровић, Д., Ћелић, И.,& Капустина, Ј.. (2013). Речник Куча. in Српски дијалектолошки зборник. Београд : Институт за српски језик САНУ. 60.