Dialectometric study of Udmurt varieties

Timofey Arkhangelskiy

Universität Hamburg

The work is supported by RFBR grant 20-512-14003 ASCF_a "Linguistic diversity in the Volga-Kama region. Typology and language documentation between Volga and Urals"

Dialectology

- Uses linguistic geography to study and classify areal varieties of languages
- Huge dialectological atlases of most European languages have been compiled since late 19th century
- Traditionally, the main focus has been on phonology and vocabulary, somewhat less on morphology
- Competing dialectal classifications often exist for the same language

Dialectometry

- Uses mathematical methods and visualizations for analyzing dialectal data as represented in atlases
- Appeared in 1970s, when Séguy (1971, 1973) and Goebl (1982) applied statistical methods for studying Romance varieties
- The source data is always a table derived from an atlas
  - Rows correspond to settlements
  - Columns correspond to features / atlas maps
  - Each cell contains value(s) of a particular feature attested in a particular settlement
- A review can be found Wieling and Nerbonne (2015)
- There is also corpus dialectometry (Szmrecsanyi 2012)

Dialectometry

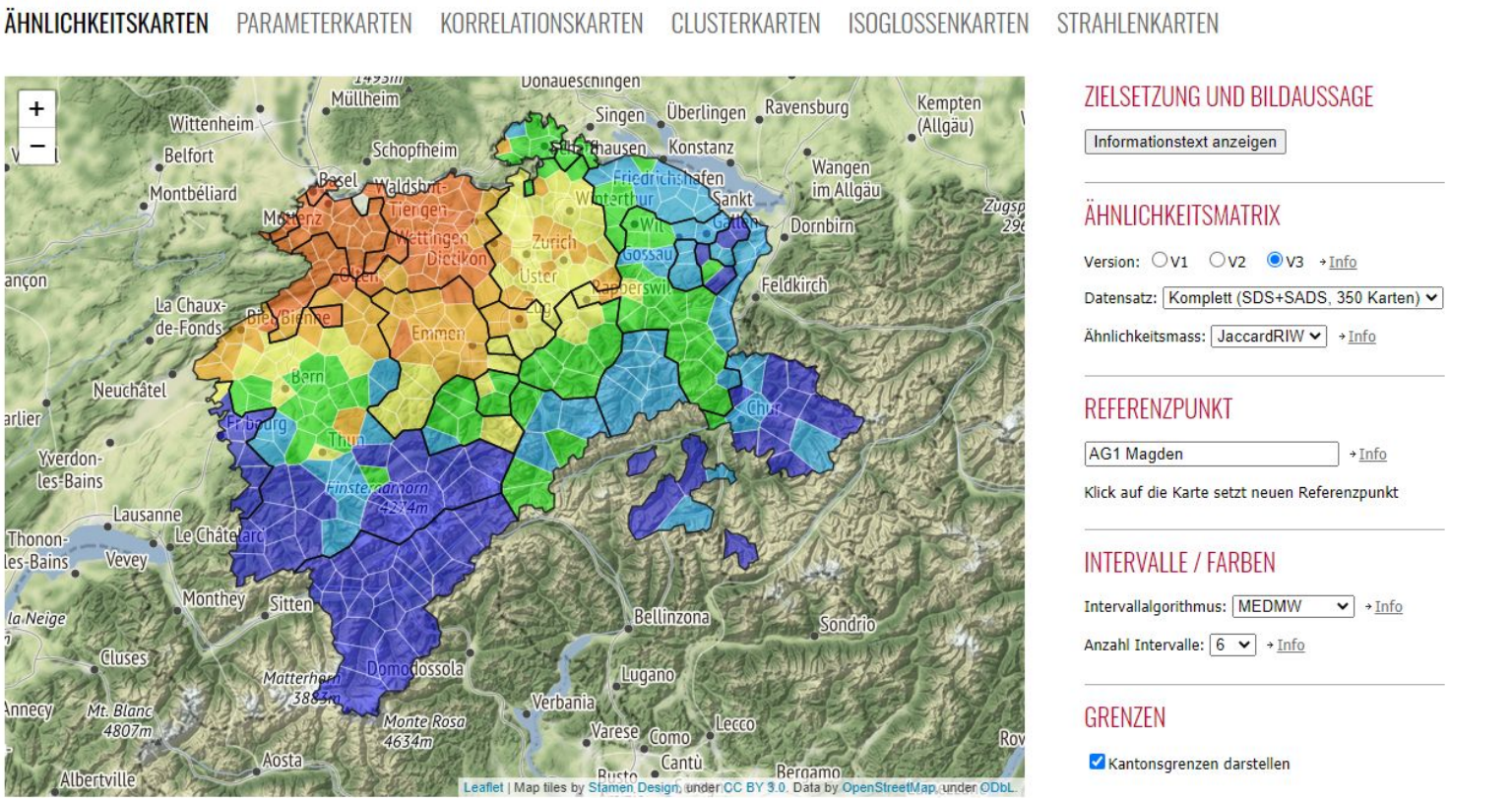
- Calculate distances between settlements using some distance function d(X, Y):
  - If all features have same values in X and Y, then d(X, Y) = 0
  - Symmetry: d(X, Y) = d(Y, X)
  - The more different values in X and Y, the greater d(X, Y)
- Several commonly used functions
  - Like Hamming distance (number of maps in which the values differ)
  - Different functions may lead to different results
  - It's good to try several functions and make sure the result is not seriously affected by function choice
  - Fortunately, this is our case

Dialectometry

- What do you do with the distance matrix?
- Classification / clusterization of varieties
  - Either hierarchical clustering; the most widely used algorithm is Ward (1963)
  - Either multidimensional scaling (Embleton 1993)
  - We focus on clustering
- Visualizations
  - Reflect clustering results in color
  - Draw isogloss bundles
  - Reveal networks of very similar varieties (continua)

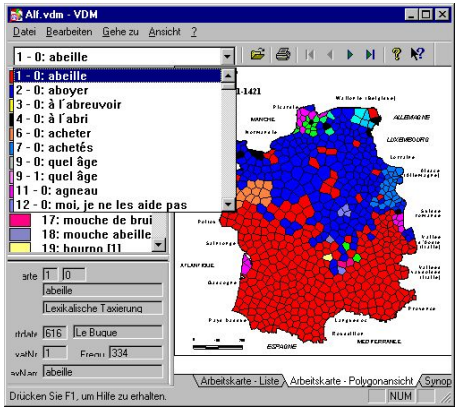
Visualization example

- http://dialektkarten.ch/ for Swiss German (Scherrer 2019):



Tools

- VDM: http://ald.sbg.ac.at/dm/Engl/VDM/features.htm (Edgar Haimerl, Hans Goebl)



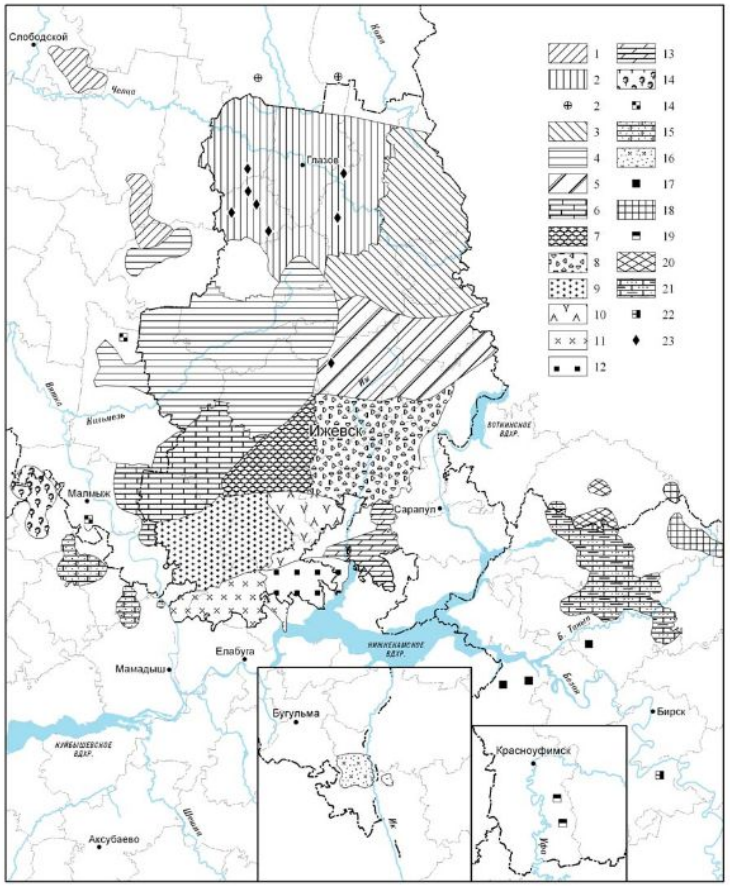
  - One of the first dialectometric tools that set the standard
- https://gabmap.nl/ (web app; Nerbonne et al. 2011)
- Aurrekoetxea (2013)
- Do-it-yourself approach

Udmurt dialectology

- Long tradition starting from Wiedemann (1858)
- Overview and bibliography can be found in Kelmakov (1998)
- Dialectological atlas (Nasibullin et al.) has been published in volumes since 2009; 7 volumes exist today
- About 20-30 maps per volume
- (Almost) each map has data from 145 settlements
- Traditional classification by Maksimov (2002: 201 and later revisions):
  - four supradialects (*наречие*): Northern, Cetral, Southern and Beserman
  - supradialect > dialect > subdialect

Udmurt dialects (Maksimov)

Распространение диалектов удмуртского языка



I. Северное наречие: 1 — нижнечепецкий диалект, 2 — среднечепецкий диалект, 3 — верхнечепецкий диалект (в т.ч. тыловайский говор).

II. Срединные говоры: 4 — средне-западный диалект (в т.ч. прикильмезские говоры), 5 — средне-восточный диалект, 6 — водзимоньинско-омгинский диалект, 7 — ныльгинский говор, 8 — среднеижский диалект.

III. Южное наречие: а) центрально-южный (собственно южный) диалект: 9 — кизнерско-можгинский говор, 10 — средне-южный говор, 11 — граховский говор, 12 — алнашский говор, 13 — кырыкмасский говор; б) периферийно-южный диалект: 14 — шошминский говор, 15 — кукморский говор, 16 — бавлинский говор, 17 — ташкичинский говор, 18 — татышлинский говор, 19 — красноуфимский говор, 20 — шагиртско-гондырский говор, 21 — буйско-танып ский говор, 22 — канлинский говор.

IV. 23 — бесермянское наречие.

Objectives

- Does it make sense to have a strict three-level classification of Udmurt dialects at all?
  - It may happen that Udmurt is a homogenous continuum, so that different isoglosses split the area in completely different ways
- If it does, would dialectometric classification coincide with the traditional one?

Data

- 42 maps from the first three volumes of the atlas
  - The results didn’t change much when adding more maps
- Ignore phonetic and other minor differences; only encode different lexemes differently
- If no value is present for a settlement, extrapolate it from the nearest points
- If there are multiple variants for a settlement, list them all as unordered set

Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	#region	#district	#settlement_rus	#settlement_udm	#number	#y	#yes	#apple	#forest	#carrot	#cucumber	#cockroach	#cat	#turkey	#moustache	#saw	#street	#window
2	Кировская область	Слободской	Нижнее Могачино	Понул	1	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй	индюк	ус	пила	ульча	косяк
3	Кировская область	Слободской	Светозарево	Гырдор	2	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй	индюк	ус	пила	ульча	косяк
4	Кировская область	Слободской	Сизёво	Бигра	3	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй	шора	ус	пила	ульча	косяк
5	Кировская область	Омутнинский	Малая Малаговская	Малая Малаговская	4	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй/кочӵыш	шора	ус?	пила	ульча	косяк
6	Кировская область	Афанасьевский	Лытка	Лытка	5	ъ/ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй/кочӵыш	курка/немыч	ус	пила	ульча	косяк
7	Кировская область	Зуевский	Слудка	Гордъяр	6	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	кочӵыш	шора	ус	пила	ульча	косяк
8	Кировская область	Фалёнский	Левановo	Левановo	7	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй/кочӵыш	шора?	ус	пила	ульча	косяк
9	Кировская область	Фалёнский	Баженово	Бажгурт	8	ы	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	кочӵыш	шора	ус	пила	ульча	косяк
10	Кировская область	Унинский	Астрахань	Асьтаркан	9	ы	бен	яблок	ӵашша	ӵужкушман	огреч	таракан	кочӵыш	шора	ус	пила	ульча	косяк
11	Кировская область	Унинский	Удмуртский Сурвай	Сырвай	10	ы	да	яблок	тэль	ӵужкушман	огреч	таракан	кочӵыш	шора	ус	пила	ульча	укно
12	Кировская область	Унинский	Малый Полом	Пичи Полом	11	ы	да	яблок/пуяблок	тэль	ӵужкушман	огреч	таракан	кочӵыш	шора	ус	пила	ульча	укно
13	Кировская область	Кильмезский	Кульма	Кудма	12	ы	0	яблок	тэль	кӟшыр	кияр	таракан	кочӵыш	немыч	ус	пила	урам	укно
14	Кировская область	Кильмезский	Салья	Салья	13	ы	0	яблок	тэль	кӟшыр	кияр	таракан	кочӵыш	немыч	ус	пила	урам	укно
15	Кировская область	Кильмезский	Паска	Поска	14	ы	0	яблок	тэль	ӵужкушман	кияр	таракан	кочӵыш	шора	ус/туш	пила	урам	укно
16	Кировская область	Малмыжский	Порез	Порез	15	ы	бен	улмо	нюлэс	кӟшыр	кияр	таракан	кочӵыш	немыч	мыйык	бычкы	урам	укно
17	Кировская область	Малмыжский	Большая Шабанка	Шубон	16	ы	0	яблок	тэль	кӟшыр	кияр	таракан	кочӵыш	немыч	мыйык	пила	урам	укно
18	Кировская область	Малмыжский	Удмуртский Китяк	Кетэг	17	ы	бен	улмо	нюлэс	кӟшыр	кияр	таракан	кочӵыш	немыч?	мыйык	бычкы	урам	укно
19	Кировская область	Вятско-Полянский	Дым-Дым-Омга	Камдор-Омга	18	ы	0	яблок	тэль	ӵужкушман	кияр	таракан	кочӵыш	немыч	мыйык	пила	урам	укно
20	Удмуртия	Ярский	Бозино	Бозьпи	19	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй/кочӵыш	шора	ус	пила	ульча	косяк
21	Удмуртия	Ярский	Тум	Дӧжигурт	20	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй/кочӵыш	шора/курка?	ус	пила	ульча	косяк
22	Удмуртия	Ярский	Зюино	Жуйна	21	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй	шора	ус	пила	ульча	косяк
23	Удмуртия	Ярский	Укан	Укана	22	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй	шора	ус	пила	ульча	косяк
24	Удмуртия	Ярский	Никольское	Никольск	23	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй	шора	ус	пила	ульча	косяк
25	Удмуртия	Глазовский	Золотарёво	Пӧжы	24	ы^/ы	бен	яблок	ӵашша	морков/ӵужкушман	огреч	торокан	писяй/кочӵыш	курка	ус	пила	ульча	косяк
26	Удмуртия	Глазовский	Дондыкар	Дондыкар	25	ъ/ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка	ус	пила	ульча	косяк
27	Удмуртия	Глазовский	Люм	Люм	26	ъ	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка/немыч	ус	пила	ульча	косяк
28	Удмуртия	Глазовский	Курегово	Куреггурт	27	ы	бен	яблок	ӵашша/нюлэс	ӵужкушман/морков	огреч	торокан	писяй/кочӵыш	немыч	ус	пила	ульча	косяк
29	Удмуртия	Глазовский	Кожиль	Нюры	28	ъ/ы	бен	яблок	ӵашша	морков/ӵужкушман	огреч	торокан	писяй	шора	ус	пила	ульча	косяк
30	Удмуртия	Глазовский	Штанигурт	Штанигурт	29	ъ	бен	яблок	ӵашша	морков/ӵужкушман	огреч	торокан	писяй	курка/немыч/шора	ус	пила	ульча	косяк
31	Удмуртия	Глазовский	Гулёково	Гылёгурт	30	ъ	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка/шора	ус	пила	ульча	косяк
32	Удмуртия	Глазовский	Пусошур	Шӧмпи	31	ы^	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка	ус	пила	ульча	косяк
33	Удмуртия	Балезинский	Верх-Люкино	Родькагурт	32	ы	бен	яблок	нюлэс	ӵужкушман/морков	огреч	торокан	писяй	курка	ус	пила	ульча	косяк
34	Удмуртия	Балезинский	Оросово	Оросгурт	33	ы	бен	яблок	нюлэс/ӵашша	ӵужкушман/морков	огреч	торокан	писяй/кочӵыш	курка	ус	пила	ульча	косяк
35	Удмуртия	Балезинский	Большой Унтем	Чабъя	34	ы	бен	яблок	нюлэс	ӵужкушман/морков	огреч	торокан	писяй	курка/немыч	ус	пила	ульча	косяк
36	Удмуртия	Балезинский	Пыбья	Пӧбъя	35	ъ/ы	бен	яблок	нюлэс/ӵашша	морков/ӵужкушман	огреч	торокан	писяй	курка/немыч	ус/мыйык	пила	ульча	косяк
37	Удмуртия	Балезинский	Юнда	Юнды	36	ъ	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй/кочӵыш	курка	ус	пила	ульча	косяк
38	Удмуртия	Балезинский	Андрейшур	Андрейшур	37	ы	бен	яблок	сик	ӵужкушман	огреч	торокан	писяй/кочӵыш	курка	ус/мыйык	пила	ульча	косяк/укно
39	Удмуртия	Кезский	Старая Гыя	Вуж Гойы	38	ы	бен	яблок	нюлэс	морков/ӵужкушман	огреч	торокан	писяй/кочӵыш	немыч	ус	пила	ульча	косяк
40	Удмуртия	Кезский	Новый Унтем	Выльгурт/Унтэм	39	ы	бен	яблок	нюлэс	морков/ӵужкушман	огреч	торокан	кочӵыш	немыч	ус	пила	ульча	косяк
41	Удмуртия	Кезский	Александрово	Кузяммувыр	40	ы	бен	яблок	нюлэс	ӵужкушман	огреч	торокан	писяй/кочӵыш	немыч	ус	пила	ульча	косяк
42	Удмуртия	Кезский	Юски	Югурт	41	ы	бен	яблок	нюлэс	ӵужкушман	огреч	торокан	кочӵыш	немыч	ус	пила	ульча	косяк
43	Удмуртия	Кезский	Пужмезь	Пужмезь	42	ы	бен	яблок	нюлэс	ӵужкушман	огреч	торокан	писяй/кочӵыш	индюк	ус	пила	ульча	косяк
44	Удмуртия	Кезский	Удмурт-Зязьгор	Зязьгор	43	ы	бен	яблок	нюлэс/сик	ӵужкушман	огреч	торокан	кочӵыш	немыч	ус	пила	ульча	косяк
45	Удмуртия	Кезский	Полом	Ойыл	44	ы	бен	яблок	сик	морков/ӵужкушман	огреч	торокан	кочӵыш	курка/немыч	ус	пила	ульча	косяк
46	Удмуртия	Юкаменский	Шамардан	Шамардан	45	ъ	бен	яблок	ӵашша	морков/ӵужкушман	огреч	торокан	писяй	курка	ус	пила	ульча	косяк
47	Удмуртия	Юкаменский	Филимоново	Кионгурт	46	ъ	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй	курка	мыйык	пила	ульча	косяк
48	Удмуртия	Юкаменский	Чурашур	Чурашур	47	ъ	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка/шора	ус	пила	ульча	косяк
49	Удмуртия	Юкаменский	Верх-Уни	Чуривыл	48	ъ	бен	яблок	ӵашша	морков	огреч	торокан	писяй/кочӵыш	курка/шора	ус/мыйык	пила	ульча	косяк
50	Удмуртия	Юкаменский	Жувам	Жуам	49	ъ	бен	яблок	ӵашша	ӵужкушман	огреч	торокан	писяй	курка	ус/мыйык	пила	ульча	косяк
51	Удмуртия	Красногорский	Дёбы	Дьӧбо	50	ъ/ы^/ы	бен	яблок	ӵашша	морков/ӵужкушман	огреч	торокан	писяй	курка	ус	пила	ульча	косяк
52	Удмуртия	Красногорский	Прохорово	Писяйгурт	51	ы	бен	яблок	ӵашша	морков	огреч	торокан	писяй	курка/индюк	ус	пила	ульча	косяк

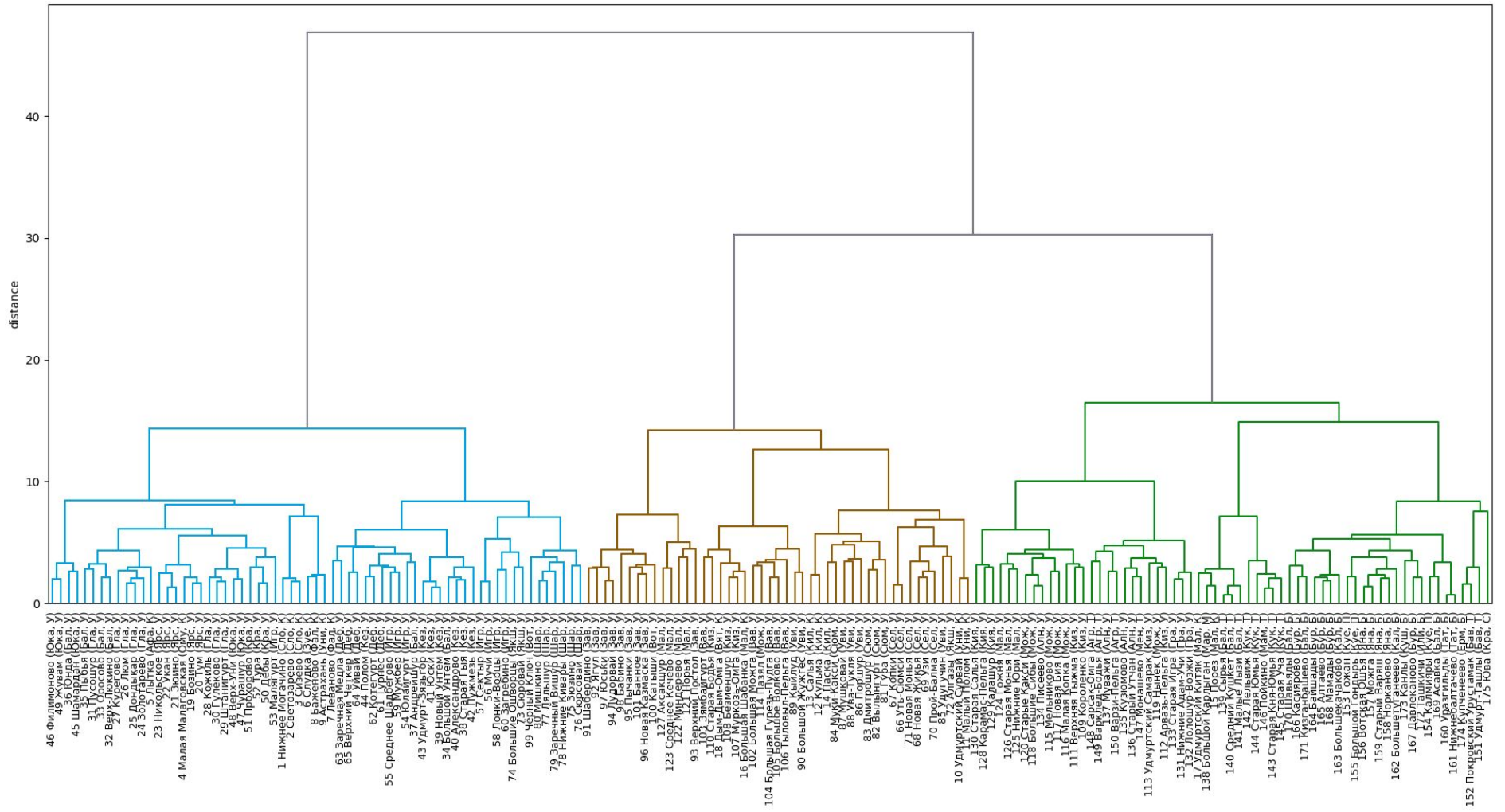
Calculations

- Euclidean distance with (almost) one-hot feature encoding
- E.g. consider three settlements:

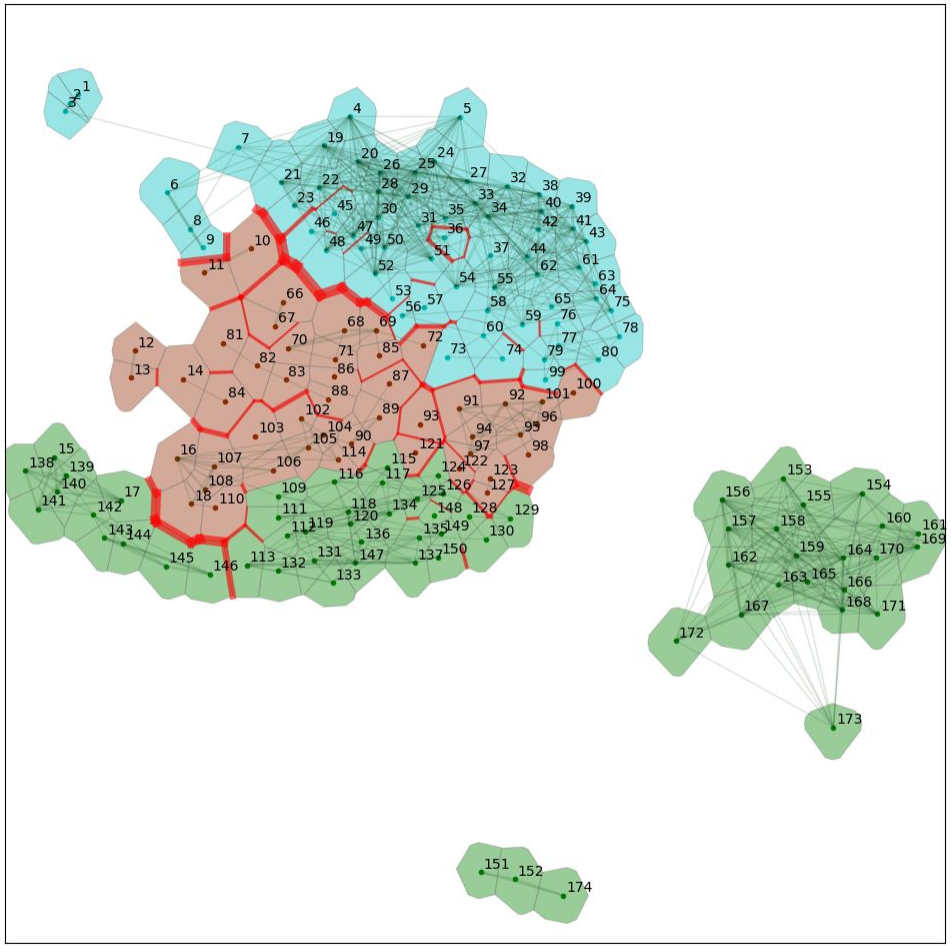
settlement	cucumber	apple
A	ogreč	jablok
B	ogreč/kijar	ulmo
C	kijar	ulmo

- Encoding: A `(1, 0, 1, 0)`, B `(0.5, 0.5, 0, 1)`, C `(0, 1, 0, 1)`
- Distances:
  - d(A, B) = 1.58
  - d(B, C) = 0.71
  - d(A, C) = 2.0

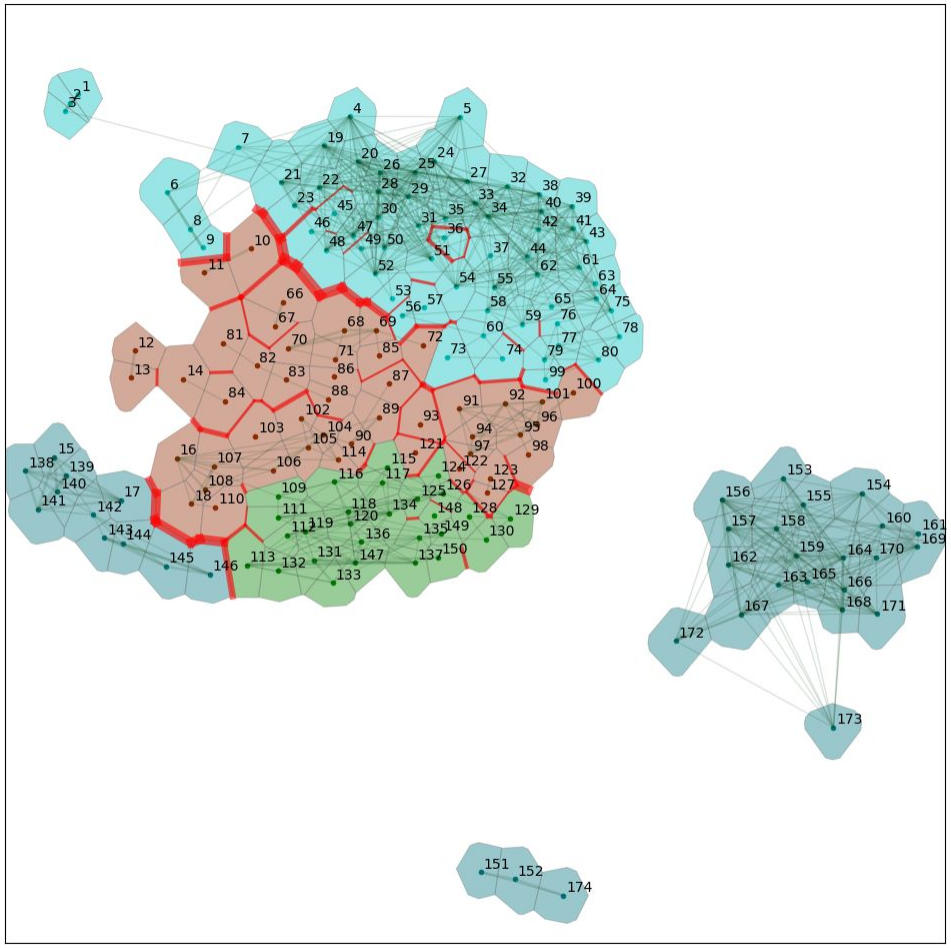
Hierarchical clusterization



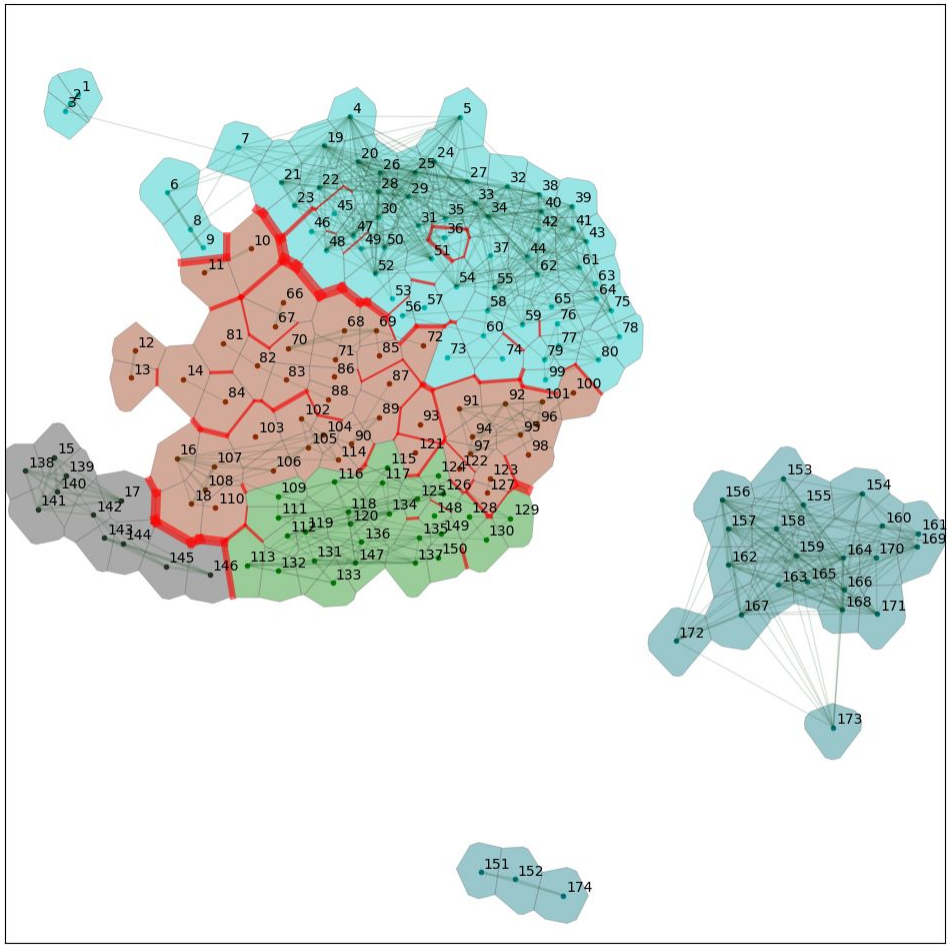
Visualization (3 clusters)



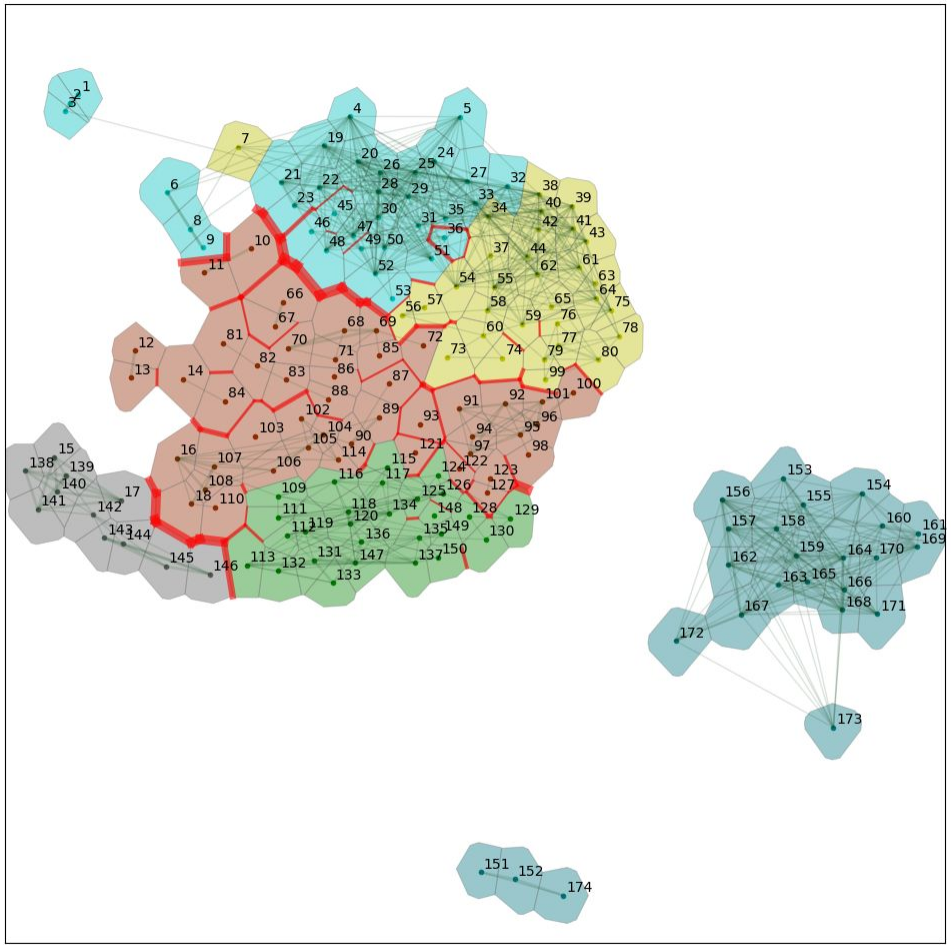
Visualization (4 clusters)



Visualization (5 clusters)



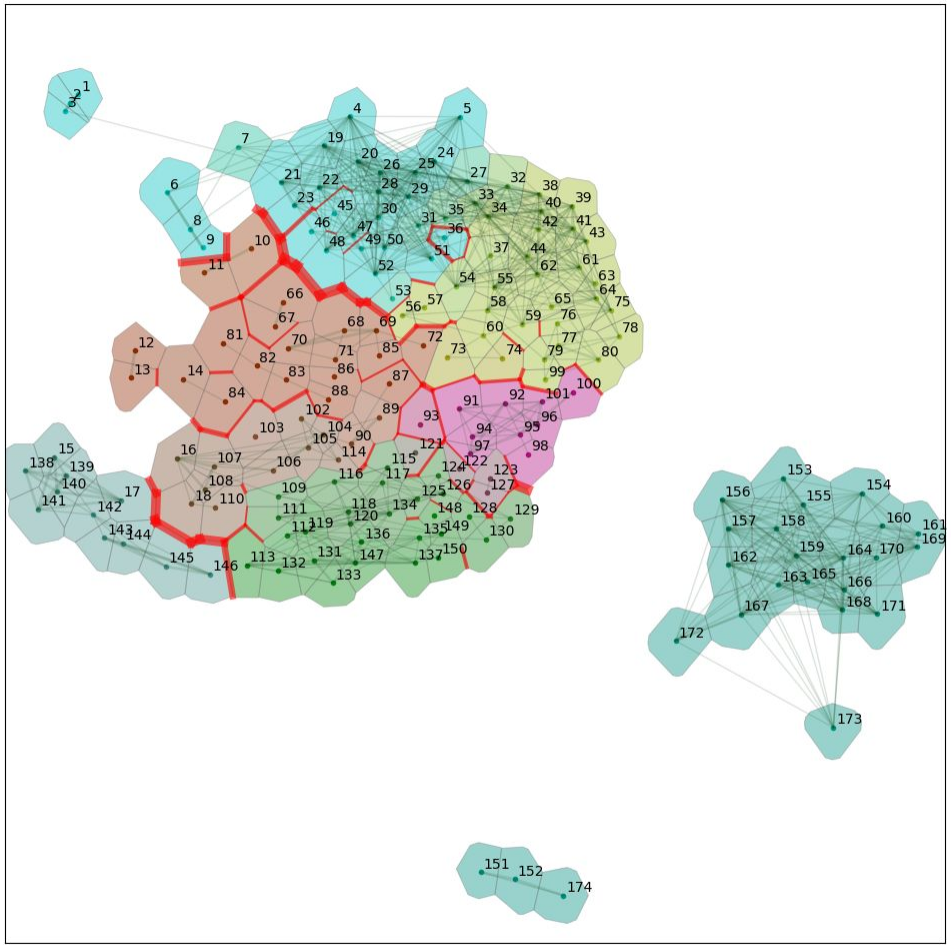
Visualization (6 clusters)



Averaging

- How robust is this classification?
- I.e. will it be the same if we use a different set of maps?
- Averaging:
  - Remove 30% of maps at random
  - Clusterize
  - Repeat 30 times
  - Visualize the result, mixing colors from the 30 iterations

Visualization (average, 7-8 clusters)



Analysis

- 3 supradialects are clearly separated, the split matches the traditional one with a couple of exceptions
  - Beserman, not so much – but there are other reasons for singling it out
- Some dialects / areas (e.g. South Peripheral) are clearly separable
- Some not
  - Much of the Northern supradialect is a very homogenous area that is hard to split into pieces
  - If you try and split it, the border depends a lot on the particular map set you choose

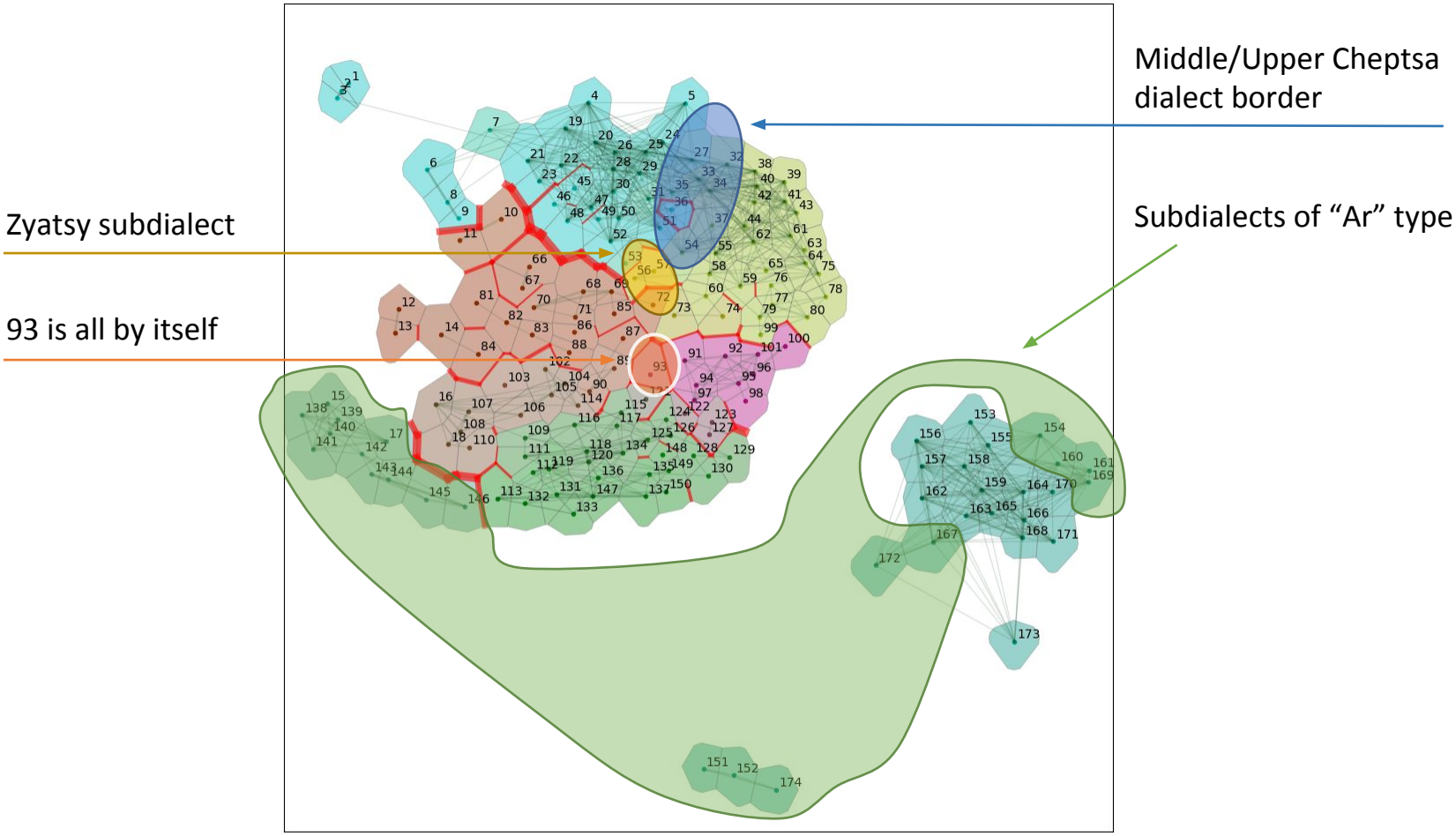
Differences

- Two possible explanation for the differences compared to the traditional classification:
- Either phonological (Ph) + morphological (M) features result in different clustering than lexical (L)
  - Results from other languages are somewhat ambiguous
  - Catalan (Goebl 2013: 113, 115): in most places, L and Ph are pretty similar, but there is Ph isogloss bundle absent on L map
  - Swiss German (Scherrer, Stoeckle 2016: 110): Ph, M and L show high correlation, but that‘s not always the case
  - Albanian (Rusakov et al. 2018): high correlation between M and L, but borders based on L are more vague
- Or some borders were drawn by cherry-picking isoglosses and do not exist in reality

Differences

- Probably it’s a little bit of both
- Even if Ph/M vs. L distinction is responsible for most differences, there is a good reason to revisit the traditional classification once more
- Anyway, it is important to understand there can be no single one-size-fits-all classification
  - When researching an areal distribution of a linguistic phenomenon, do not expect the isoglosses to align with traditional dialectal classification

Examples of differences



Conclusion

- The high-level split into supradialects is mostly consistent with traditional classification
- Linguistic distance between neighboring varieties is different within different dialects, which you cannot deduce from the classification
- There are homogenous areas than probably cannot and shouldn’t be split into subdialects

References

- Кельмаков, В. К. 1998. Краткий курс удмуртской диалектологии. Введение. Фонетика. Морфология. Диалектные тексты. Библиография. Ижевск: Изд-во Удм. ун-та.
- Максимов, С. А. 2002. Соотношение исторических пластов лексики в удмуртских диалектах по данным I вопросника ДАУЯ // Первой удмуртской грамматике 225 лет: Сб. статей. Ижевск: УИИЯЛ.
- Насибуллин, Р. Ш., С. А. Максимов, В. Г. Семёнов, Г. В. Отставнова. 2009. Диалектологический атлас удмуртского языка. Карты и комментарии. Ижевск: НИЦ «Регулярная и хаотическая динамика».
- Насибуллин, Р. Ш., С. А. Максимов, В. Г. Семенов, Г. В. Отставнова. 2010. Диалектологический атлас удмуртского языка. Карты и комментарии. Выпуск 2. Ижевск: НИЦ «Регулярная и хаотическая динамика».
- Насибуллин, Р. Ш., С. А. Максимов, В. Г. Семенов, Л. В. Бусыгина. 2013. Диалектологический атлас удмуртского языка. Карты и комментарии. Выпуск 3. Ижевск: НИЦ «Регулярная и хаотическая динамика».
- Aurrekoetxea, G., K. Fernandez-Aguirre, J. Rubio, B. Ruiz & J. Sanchez. 2013. "DiaTech": A new tool for dialectology. *Literary and Linguistic Computing* 28(1). 23–30.
- Embleton, Sheila. 1993. Multidimensional scaling as a dialectometrical technique: outline of a research project. In Contributions to quantitative linguistics. Proceedings of the first international conference on quantitative linguistics (QUAL-ICO). Eds. Reinhard Köhler & Burghard B. Rieger. Dordrecht & Boston: Kluwer Academic Publishers. 267–276.
- Goebl, Hans. 1982. Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie. Wien: Österr. Akad. Wiss.
- Goebl, Hans. 2013. La Dialectometrització dels quatre primers volums de l'ALDC : una breu presentació. *Estudis Romànics* 35. 87–116.
- Nerbonne, J., Colen, R., Gooskens, C., Leinonen, T., & Kleiweg, P. 2011. Gabmap – A web application for dialectology. *Dialectologia*, 65-89.
- Rusakov, Alexander, Maria Morozova and Maria Ovsjannikova. 2018. Linguistic complexity and lexicon of Albanian dialects: An attempt of quantitative analysis. A paper presented at the Workshop "First step towards an interactive map of Balkan linguistic features". 26–27 November 2018, University of Zurich.
- Scherrer, Yves, and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica* 24(1). 92–125.
- Scherrer, Yves, 2019, June. News from dialektkarten.ch. 2. VerbaAlpina-Arbeitstagung, Munich, 18 June 2019.
- Séguy Jean. 1971. La relation entre la distance spatiale et la distance lexicale. Rev. Linguist. Rom. 35, p. 335–57.
- Séguy Jean. 1973. La dialectométrie dans l'atlas linguistique de la Gascogne. Rev. Linguist. Rom. 37, p. 1–24.
- Szmrecsanyi, Benedikt. 2012. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge University Press.
- Ward, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58. 236–244.
- Wiedemann, F. J. 1858. Zur Dialektenkunde der wotjakischen Sprache. *Bulletin de la Classe historico-philologique de l'Académie impériale des sciences de St.-Pétersbourg*, XV. 240–256.
- Wieling, M. and Nerbonne, J., 2015. Advances in dialectometry. Annual Review of Linguistics, Vol. 1, p. 243–264.