A corpus study of dialectal features in Selkup

Maria Brykina

The University of Hamburg, Germany

Josefina Budzisch

The University of Hamburg, Germany

Sergei V. Kovylin

Laboratory «Linguistic Platforms» Ivannikov Institute for System Programming of the RAS, Moscow

Tomsk State Pedagogical University

A corpus study of dialectal features in Selkup

- Idea: Apply the methods of corpus dialectometry to the Selkup data
- Goal: Compare the clusterization of Selkup speakers to the existing classifications of Selkup dialects
- A pilot project (with 30 speakers and 15 features)

Selkup dialects: various classifications

Northern Selkup	Central Selkup	Southern Selkup
Taz	Vakh	Middle Ob
Tolka (Larjak)	Vasjugan	Chaya
Karasino	Tym	Ket
Turukhan	Narym	Upper Ob
Baikha		Chulym (†)
Eloguj		

[Glushkov 2013, following Janurik 1978]

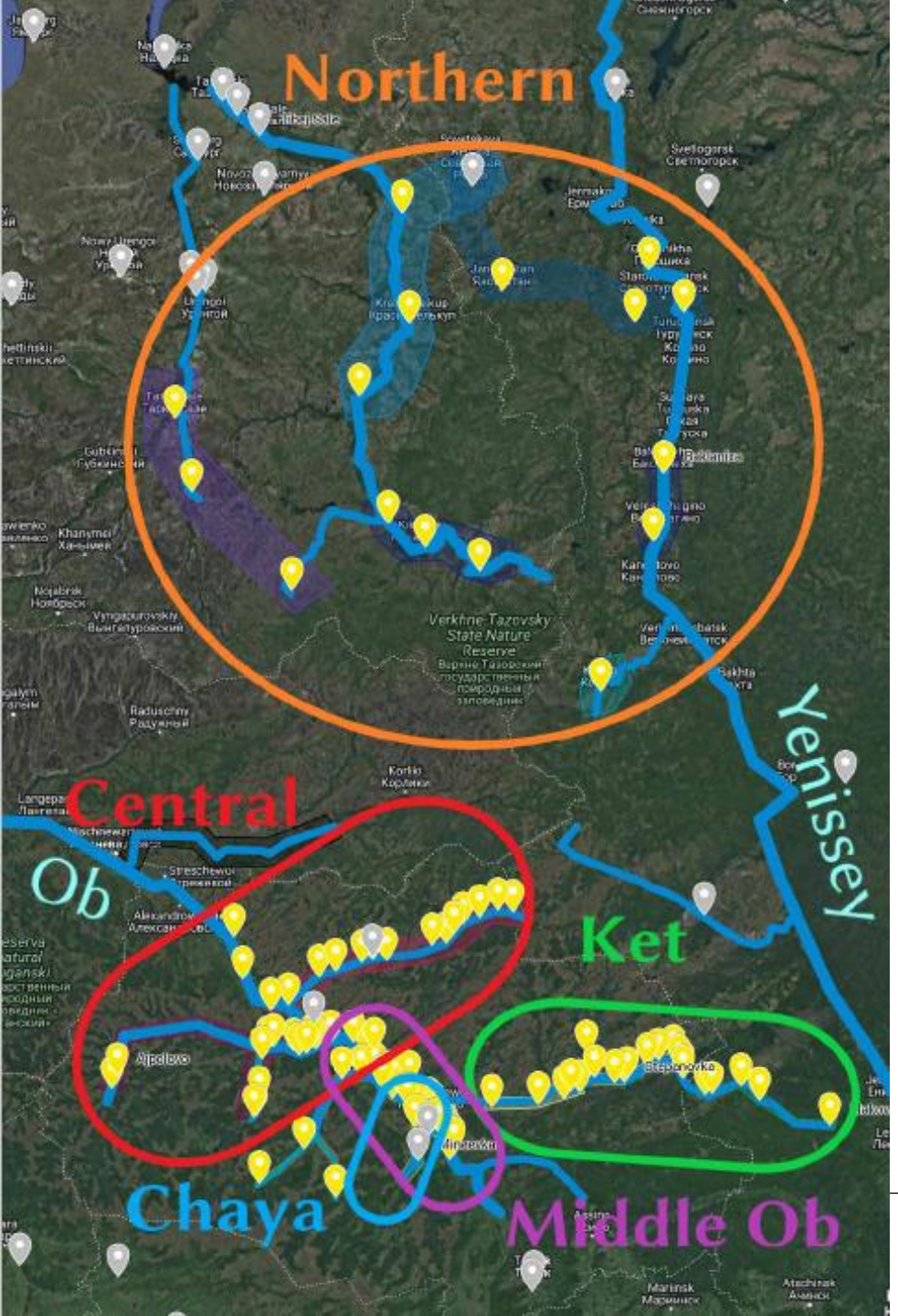
	(Central Selkup)		(Southern Selkup)	
Northern Selkup	**Tym**	**Narym**	**Ket**	**Ob**

[Helimski 2000/1985]

Data

- The INEL Selkup Corpus (https://inel.corpora.uni-hamburg.de/portal/corpora/selkup/)
  - based on the archive of Angelina Ivanovna Kuzmina
  - texts collected in 1960-1970-s from many dialects
  - about 100 speakers
- The SLC Corpus
- Selkup Corpus of Texts based on the Archive of the Laboratory of Siberian Indigenous Languages TSPU (Tomsk)
  - data collected in in 1960-1990-s from Southern and Central dialects
- Older texts

Selkup dialects in the INEL corpus



- **Northern** (> 100 speakers)
- **Central** (several (semi-)speakers)
- Southern(?): **Middle Ob, Ket** and **Chaya**

Source data: example

- Occurrences

Feature	Example	Value	NEP	YIF	KMS	TFF	PVD
Dialect			Northern	Central	Ket	MiddleOb	Chaya
P: p/m	ACC	-m	11	13	256	440	294
P: p/m	ACC	-p	143	36	0	14	0
G: DUAL	DU	Q(I)	75	4	11	22	3
G: DUAL	DU	QÄQ(I)	4	0	0	0	0
G: DUAL	DU	STJA	0	2	0	24	0
G: DUAL	DU	STJAQ(I)	0	0	0	9	13
L: say	say/tell	eǯi-	3	11	0	6	0
L: say	say/tell	čenči-	0	7	10	20	0
L: say	say/tell	kəti-	59	0	30	8	26
L: say	say/tell	t'äri-	0	0	55	0	186

Data pre-processing

- Relative frequencies within each diagnostical context (here called ‚example‘)

Feature	Example	Value	NEP	YIF	KMS	TFF	PVD
Dialect			Northern	Central	Ket	MiddleOb	Chaya
P: p/m	ACC	-m	0,07	0,27	1,00	0,97	1,00
P: p/m	ACC	-p	0,93	0,73	0,00	0,03	0,00
G: DUAL	DU	Q(I)	0,95	0,67	1,00	0,40	0,19
G: DUAL	DU	QÄQ(I)	0,05	0,00	0,00	0,00	0,00
G: DUAL	DU	STJA	0,00	0,33	0,00	0,44	0,00
G: DUAL	DU	STJAQ(I)	0,00	0,00	0,00	0,16	0,81
L: say	say/tell	eǯi-	0,05	0,61	0,00	0,18	0,00
L: say	say/tell	čenči-	0,00	0,39	0,11	0,59	0,00
L: say	say/tell	kəti-	0,95	0,00	0,32	0,24	0,12
L: say	say/tell	t’äri-	0,00	0,00	0,58	0,00	0,88

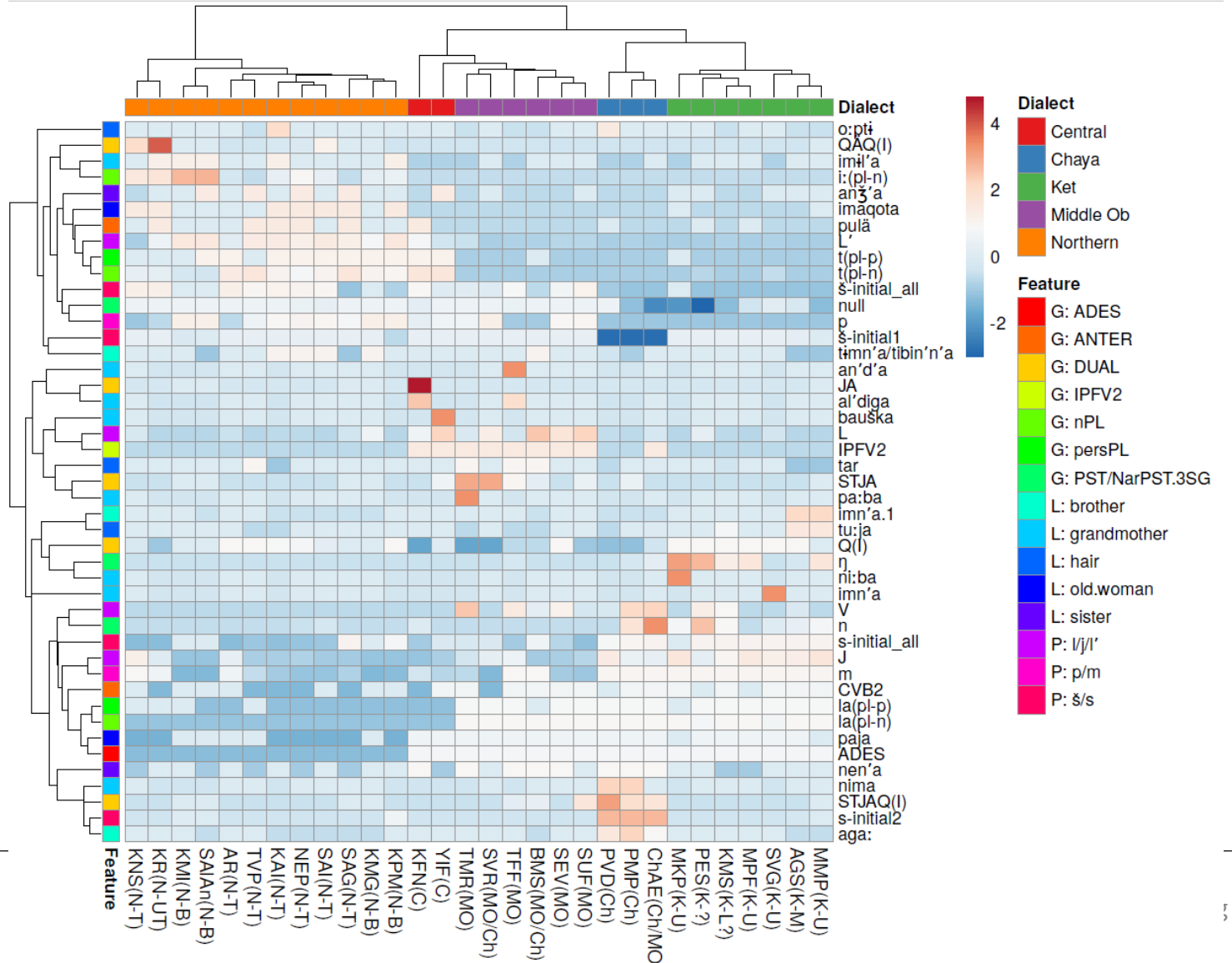
Challenges

- Corpus
  - Different amount of data (speakers and tokens) for various dialects
- Speakers
  - Different amount of data for individual speakers
  - Many speakers from whom only one text was recorded
- Missing values (→ imputation?)

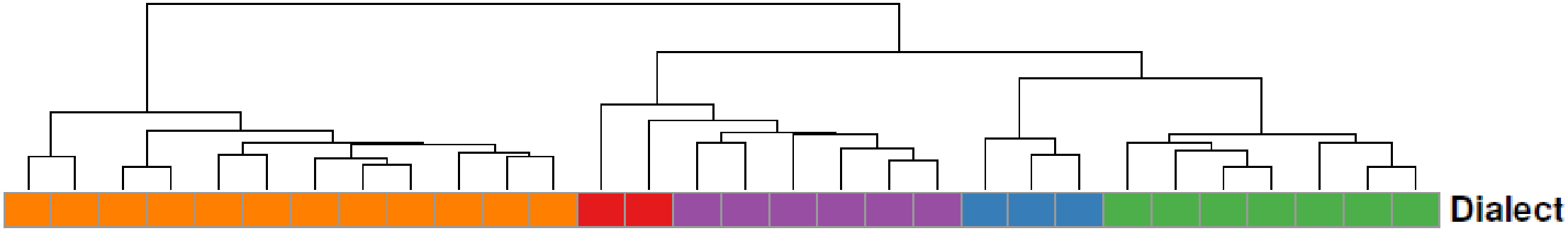
Challenges

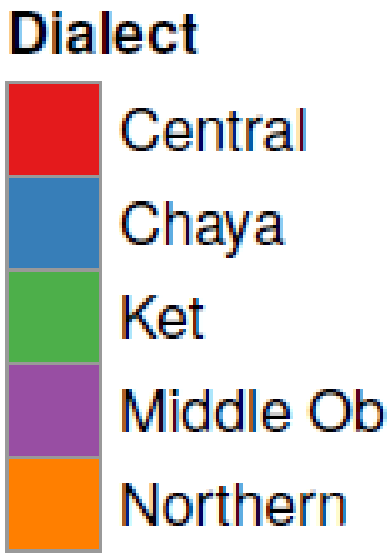
- Features
  - Features may have different number of values, which could influence the results, since features with the greater number of values will have the highest impact (→ reduce the range of each value).
  - Features may have many diagnostical contexts, which leads again to the higher impact into the overall distribution (→ calculate the average of each value through all diagnostical contexts?)
  - For now the absolute frequencies are not taken into account. Thus, we don't distinguish between 400/0 and 1/0 occurrences, both resulting after pre-processing in 1/0 values.
  - There are features attested only for some speakers (→ exclude?)
- Various methods of cluster analysis
  - For now we use Ward clusterization method and Euclidean distances (implemented in https://biit.cs.ut.ee/clustvis/)
  - Which other clusterization methods and metrics should we check?
  - Neighbor nets?

Results (only 15 features-30 speakers): clusters

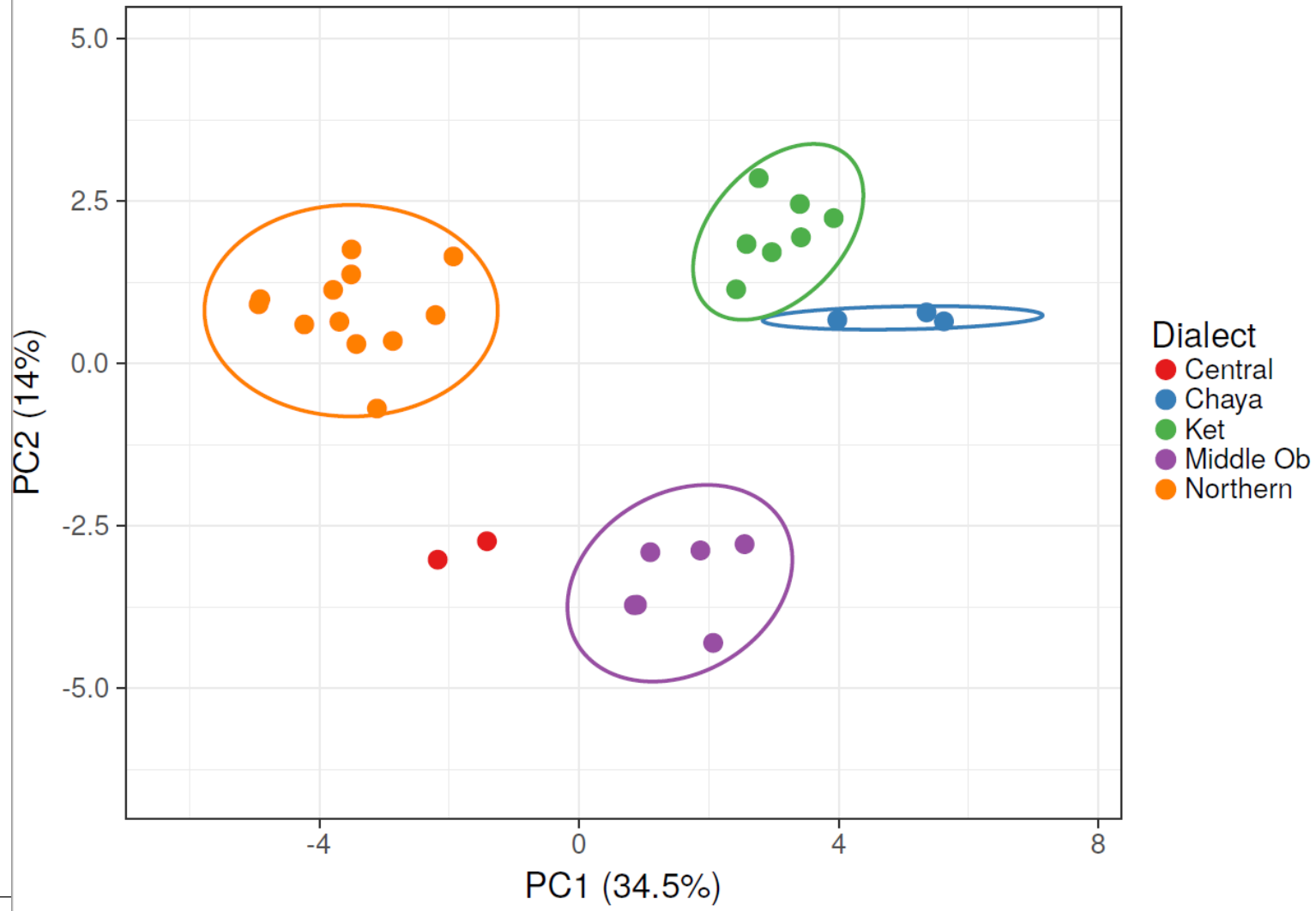


Results (only 15 features-30 speakers): clusters

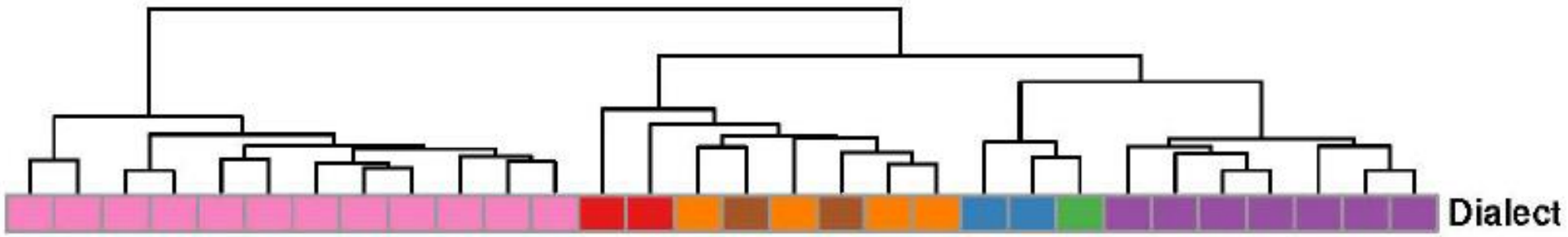


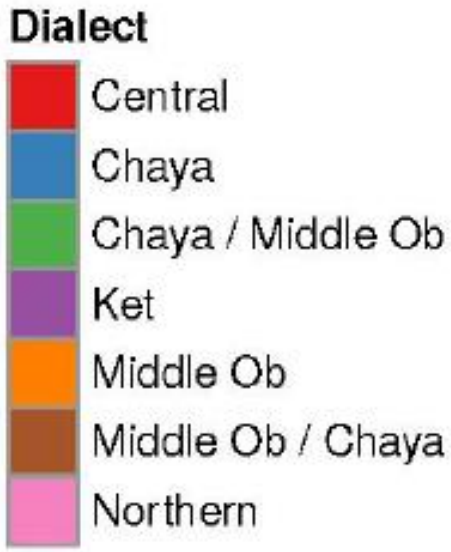


Results (only 15 features-30 speakers): PCA

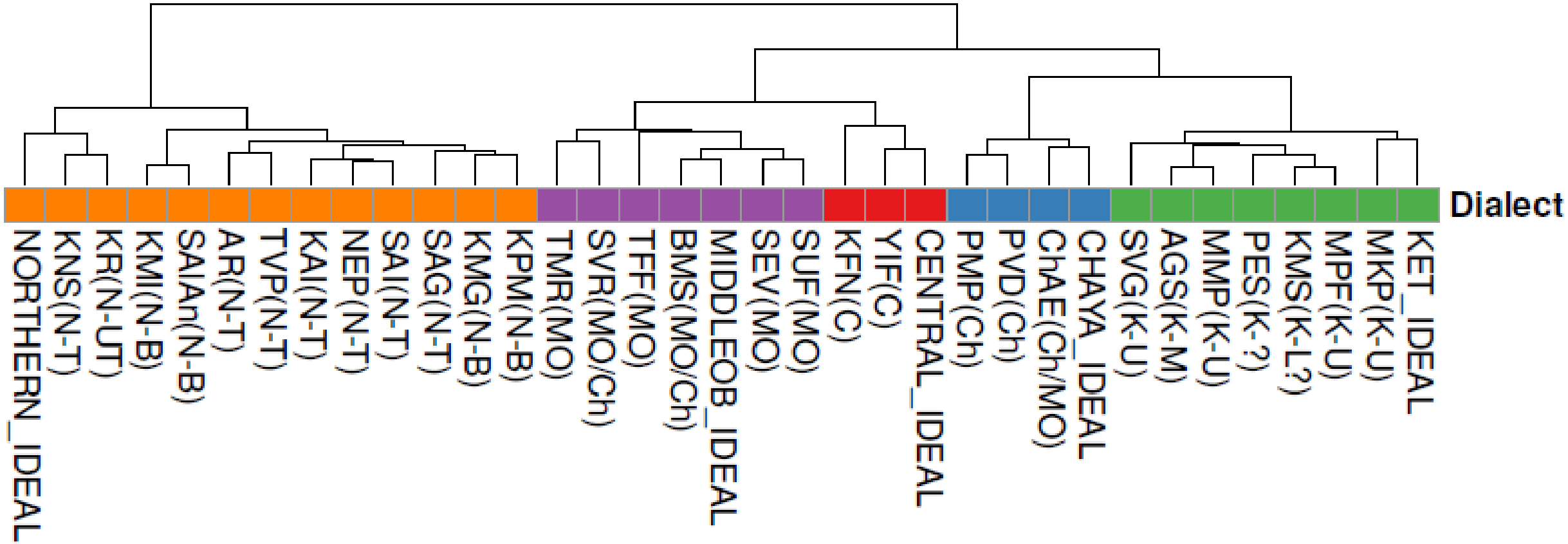


Results: dialectal attribution





Results: «ideal speakers»



Question: what to do with this data?

- It seems that such a collection of data makes sense, but what questions can we ask ourselves when we have 50 features? Will it bring us any new knowledge about the Selkup dialects?
  - Systemizing and visualizing our knowledge about features and their distribution
  - Looking at what groups we get (how the (sub-)dialects are divided into groups)?
  - What groups we get when filtering out separate features or groups of phonetic/morphological/lexical/syntactic features?
  - Attributing an unknown speaker to a dialect
  - …
  - ???

Thank you for your attention!