HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Digital Language Typology

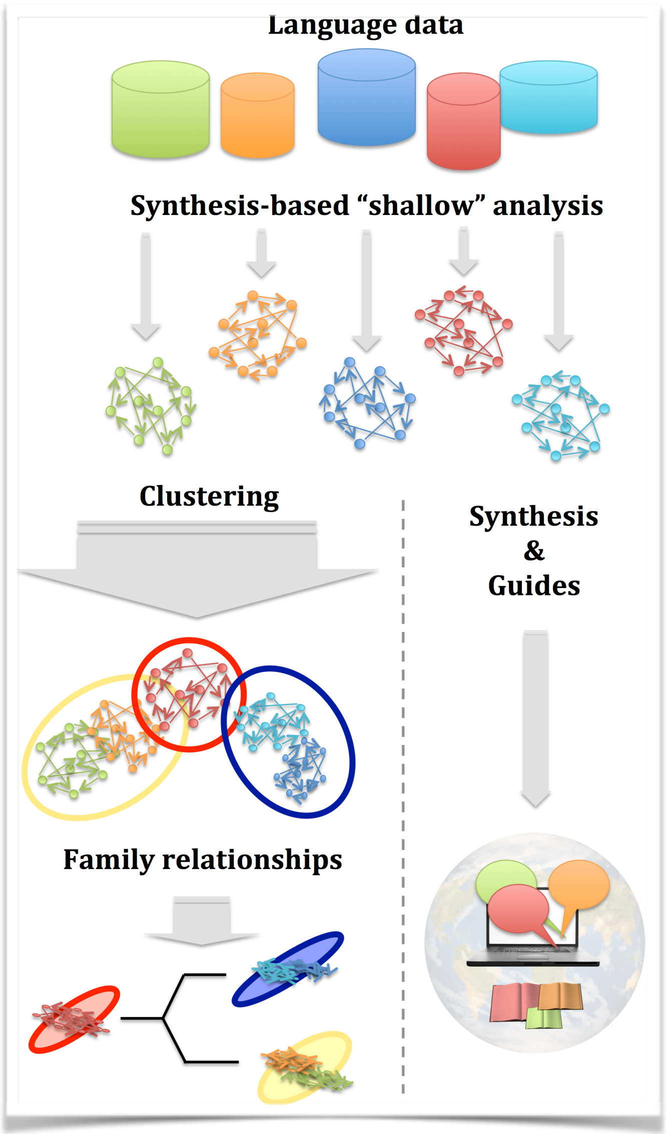
Mining from the Surface to the Core

Juraj Šimko and many others

28 November, 2017, Linguistic Convergence Laboratory Meeting, Moscow

Typology

- **Grouping** of languages according to their characteristics
- **Explaining** distributions, language contact
- **Multi-dimensional** space of similarities / differences / influence of contact: syntax, morphology, phonotactics…
- *Some* work on ***prosody*** (*Gil, 1986; Hirst & Di Cristo, 1998; Jun, 2006; Hyman, 2006; Grabe & Low, 2002*), mainly classifying languages based, e.g., on
  - lexical and postlexical intonational features
  - rhythm classes



Digital (Language Typology)

- using language/speech technology tools
- shallow, but non-trivial analysis

(Digital Language) Typology

- big, digital, language and speech data

Cummins, Gers & Schmidhuber (1999)

Automatic discrimination among languages based on prosody alone

used LSTM-based language models trained on f0 and energy contours for language comparisons based purely on these prosodic characteristics

Language n-grams and perplexity

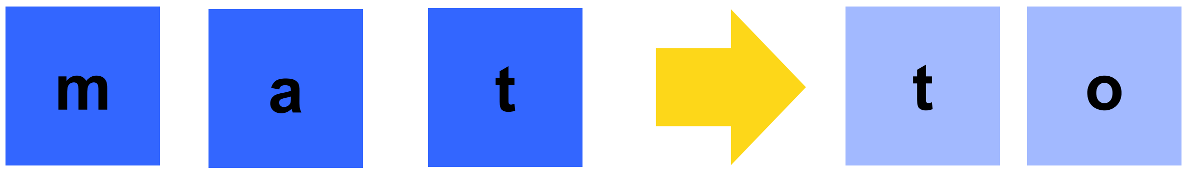
m a t → t o

$p_{\mathrm{FIN}}(\mathrm{t} \,|\, (\mathrm{t,a,m,\ldots}))$

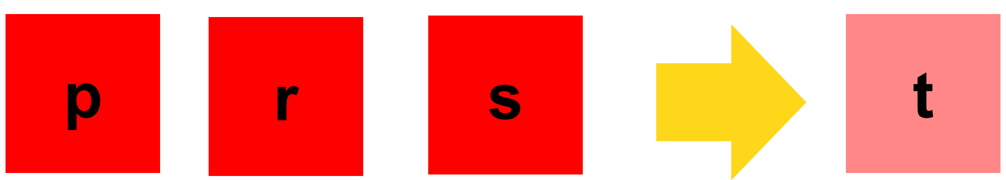
p r s → t

$p_{\mathrm{SVK}}(\mathrm{t} \,|\, (\mathrm{s,r,p,\ldots}))$

Language n-grams and perplexity



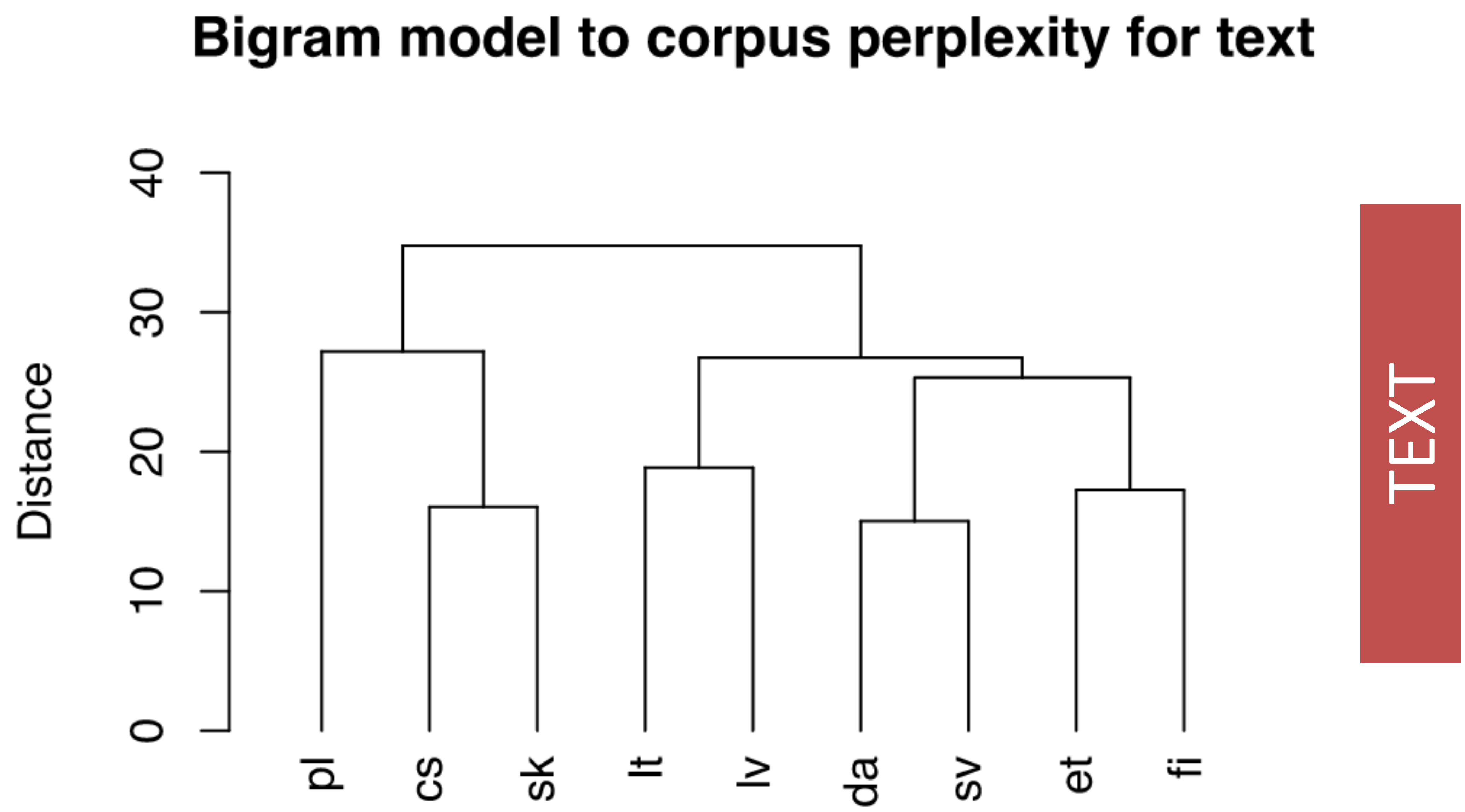
$p_{SVK}(t|(t,a,m,\ldots))$



$p_{FIN}(t|(s,r,p,\ldots))$

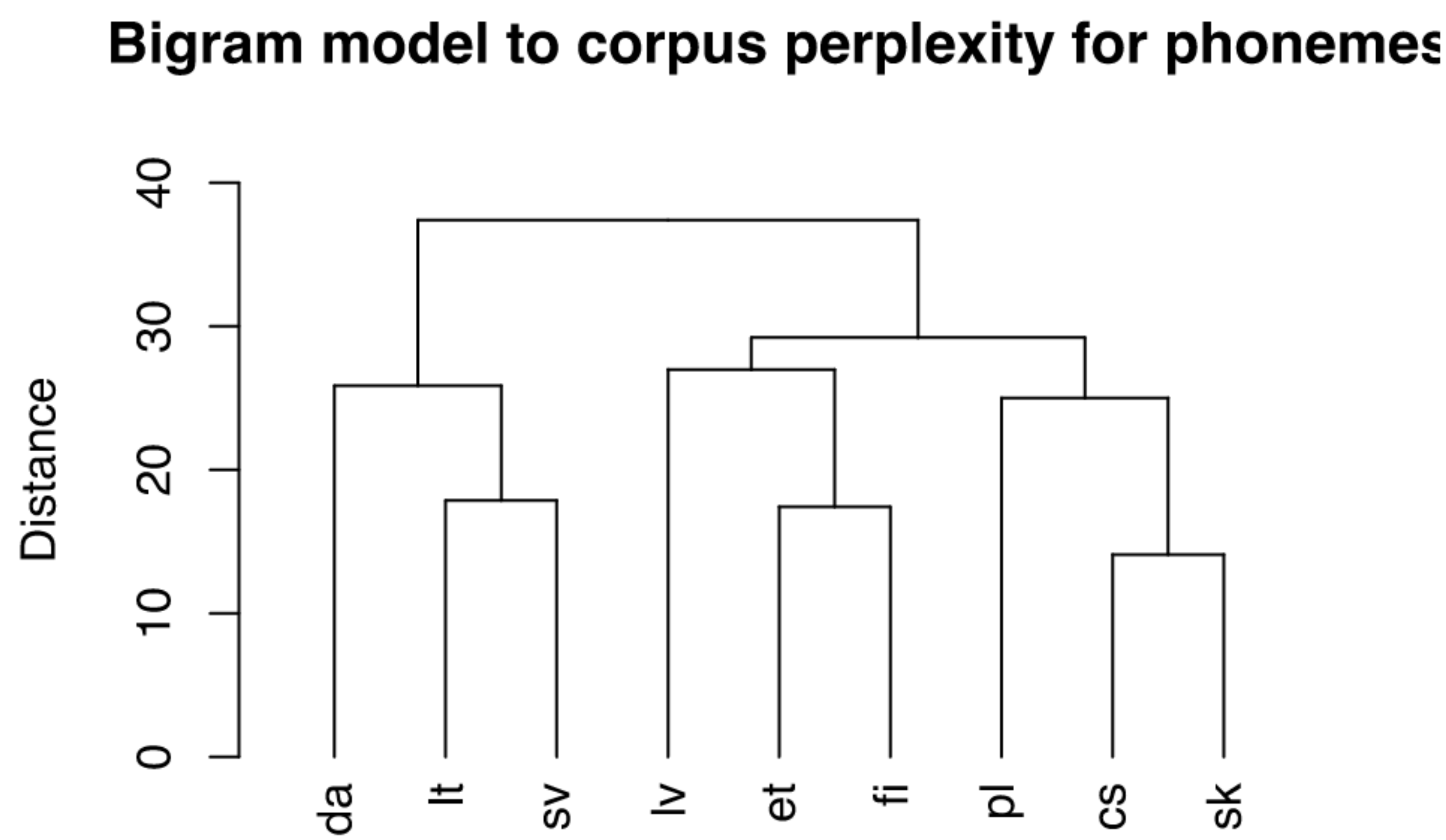
Language n-grams and perplexity

- Using the EU Europarl corpus, standard orthography



Language n-grams and perplexity

- Same corpus, transcribed using espeak



"PHONEMES"

- Not so good, non-matching phoneme sets
- We can see where the models are most perplexed: **sanity checks**

How to look at prosody?

1. Extract f_0 and energy
2. Continuous wavelet transform of the f_0 and energy signals
3. Calculate derivatives of the signals (Δ-features)
4. Discretize the Δ-feature signals: get a finite state space
5. Train simple unigram models (probabilities of individual states) for all languages separately
6. For each sentence, compute perplexity measure for each language separately
7. Using mean perplexity of a given language with sentences from all languages, create a confusion matrix
8. Plot something summarizing the confusion matrix

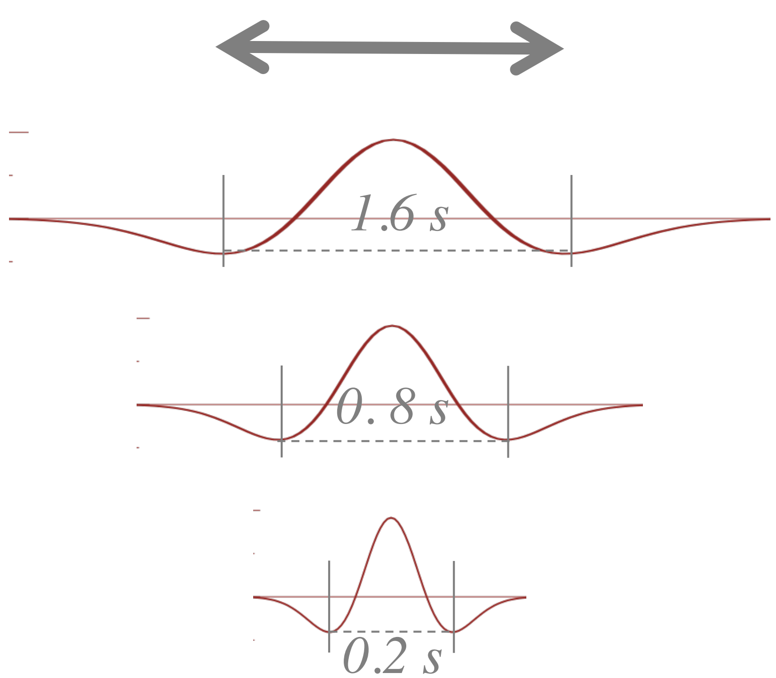
Methodology

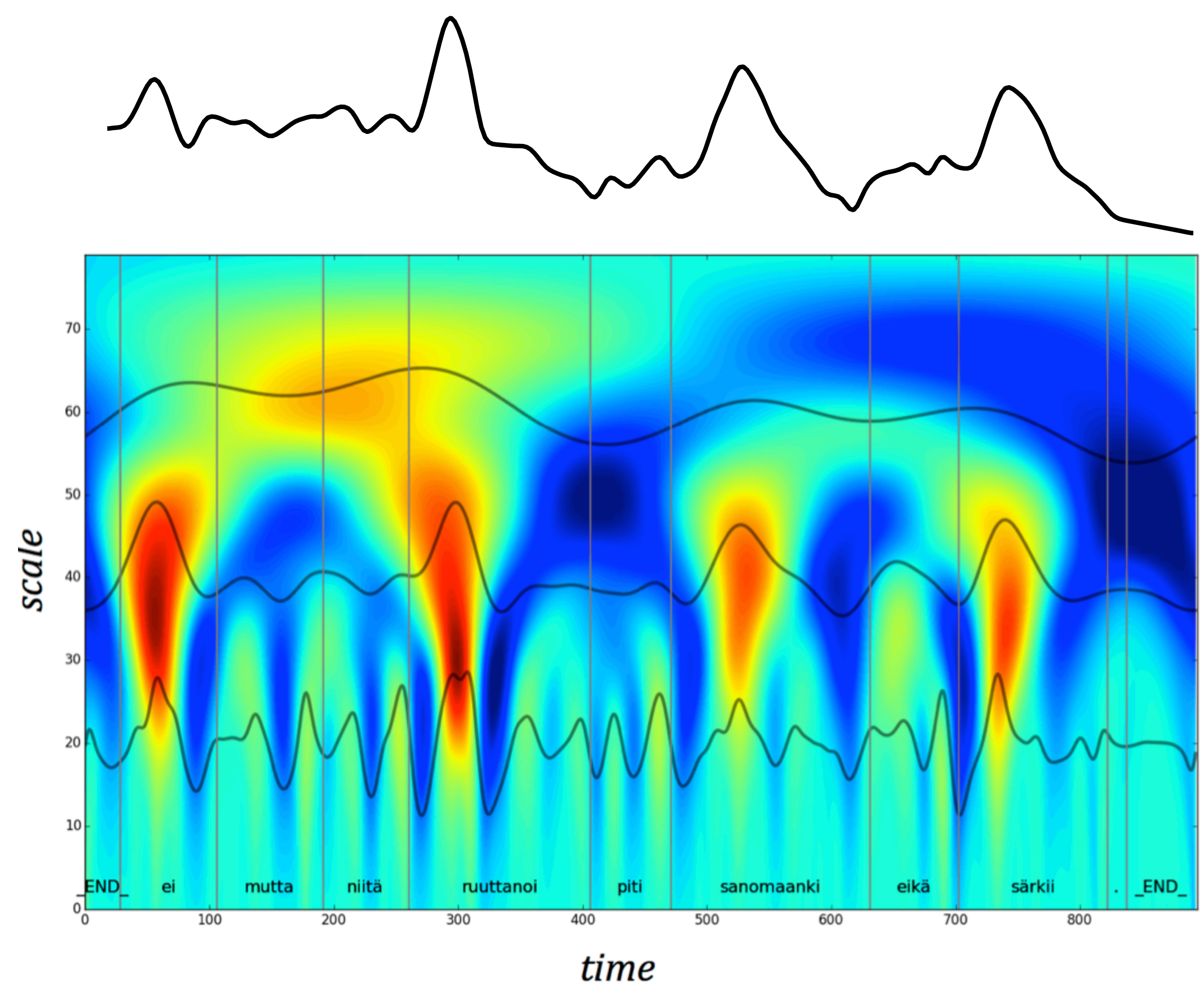
1. Extract f_0 and energy

- ✓ f_0 extracted using praat, (linearly) interpolated and smoothed (10 Hz bandwidth)
- ✓ signal envelopes (energy) contours extracted using continuous wavelet transform method (see the next slide)
- ✓ both signals sampled at 100 Hz and time-aligned

Methodology

2. Continuous wavelet transform of the f_0 and energy signals

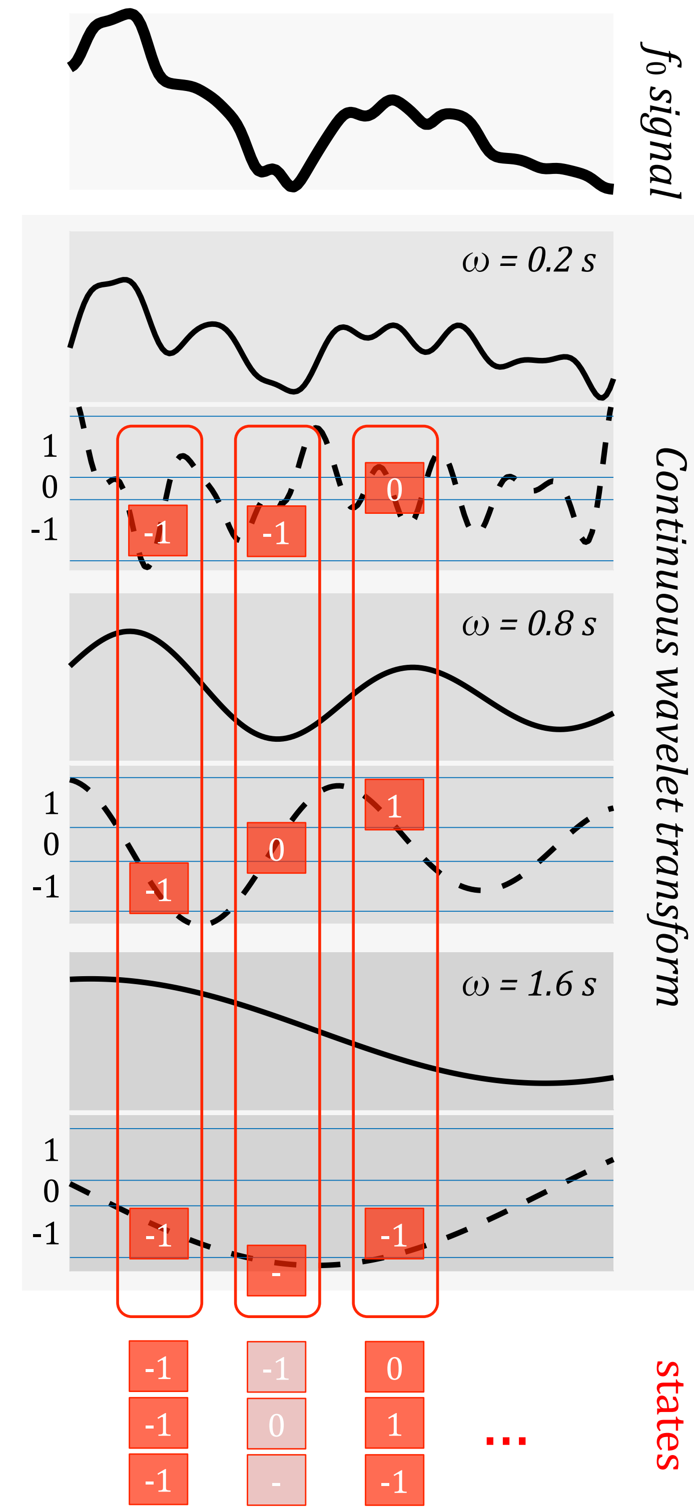




Methodology

3. Calculate derivatives of the signals (Δ-features)

4. Discretize the Δ-feature signals: get a finite state space



Methodology

5. Train simple unigram models (probabilities of individual states) for all languages separately

for each state S, compute

$$P_{\text{SWE}}(S), P_{\text{GER}}(S), P_{\text{RUS}}(S), P_{\text{SVK}}(S), P_{\text{HUN}}(S), P_{\text{EST}}(S), P_{\text{FIN}}(S)$$

6. For each sentence, compute perplexity measure for each language separately

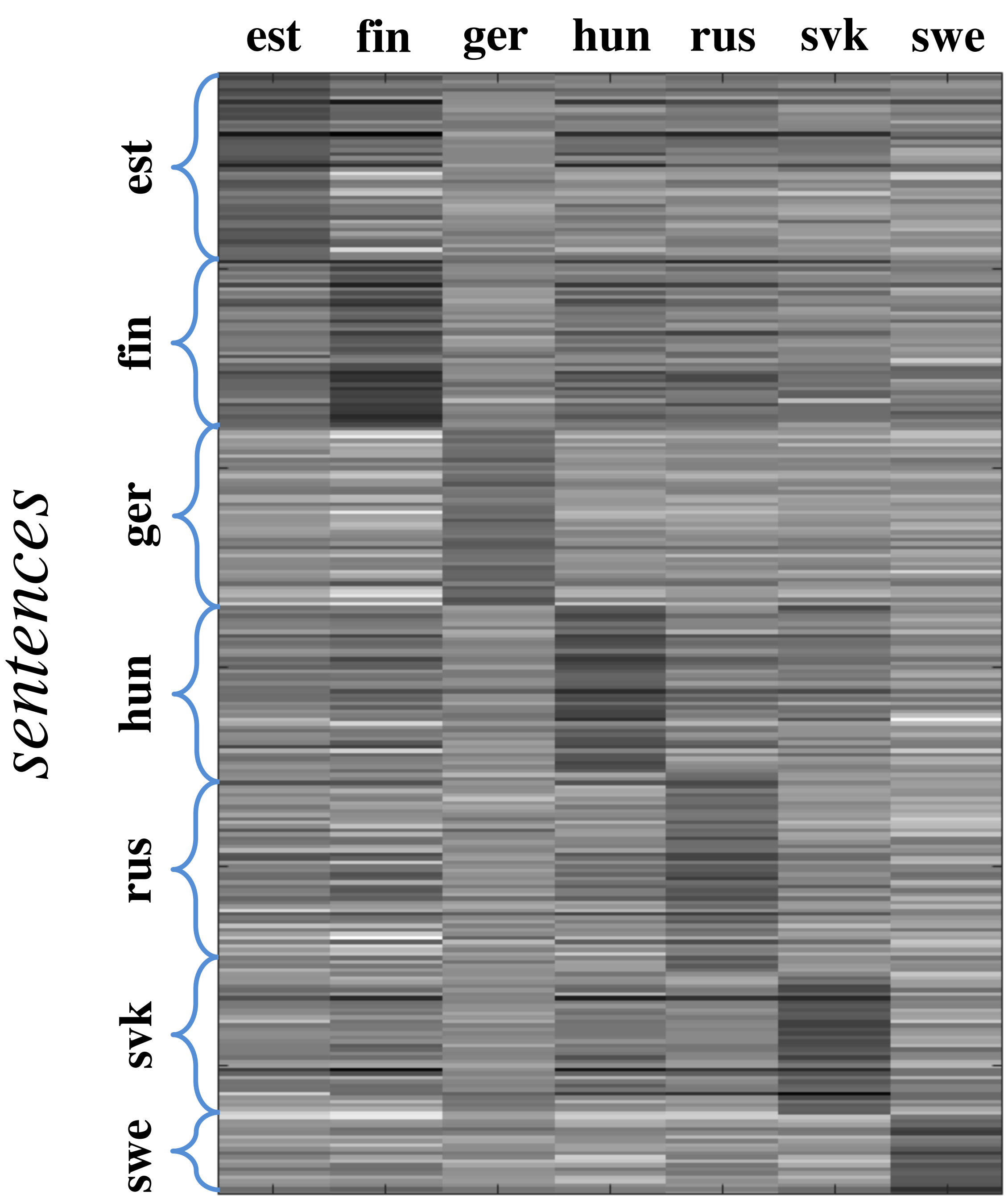
formally, for sentence $S_1 S_2 S_3 \ldots S_N$ and language LAN, perplexity is:

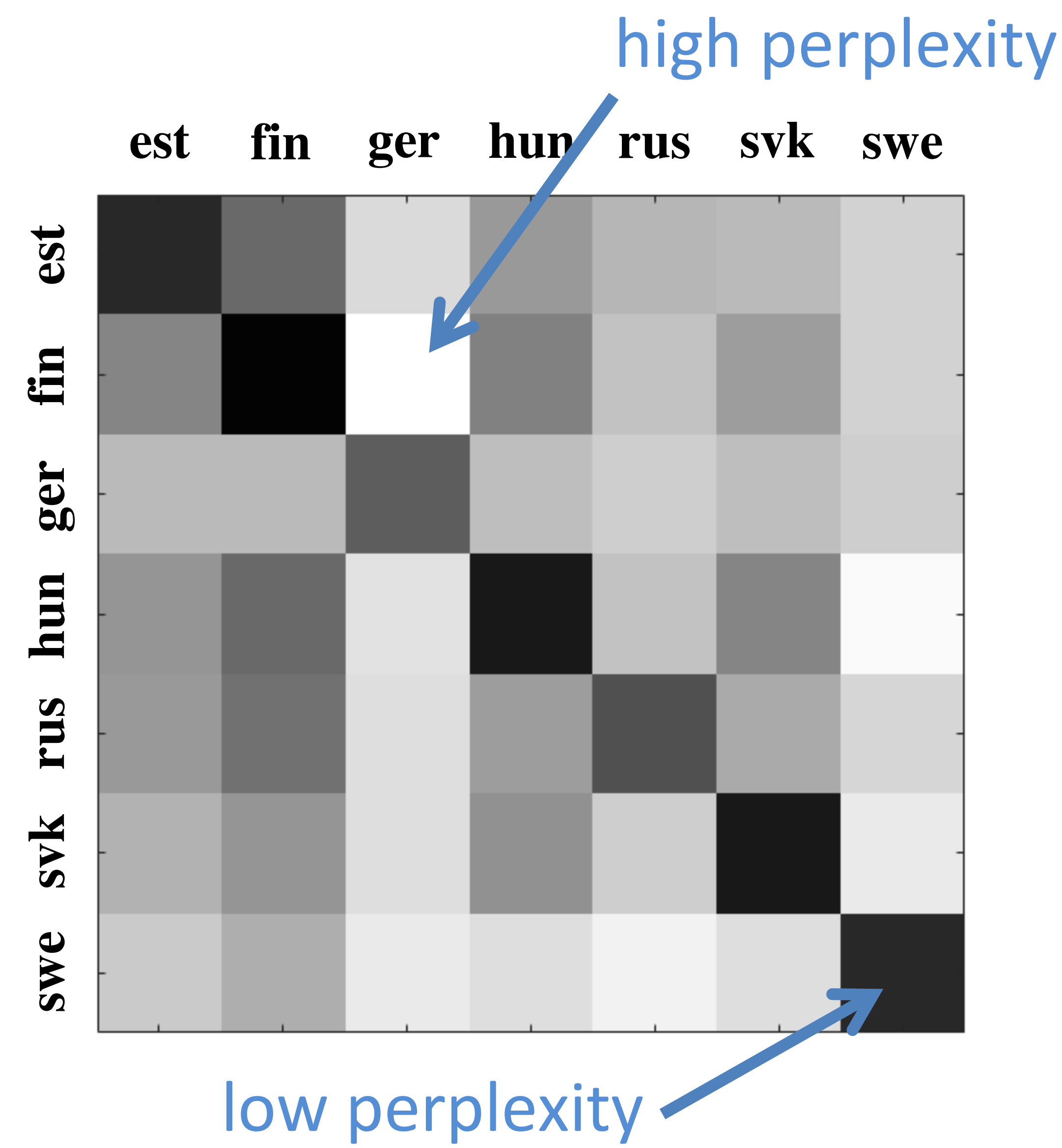
$$2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 P_{\text{LAN}}(S_i)}$$

informally, perplexity is a measure of “surprise” that the given state is found in the given sentence in the given language

Methodology

7. Using mean perplexity of a given language with sentences from all languages, create a confusion matrix

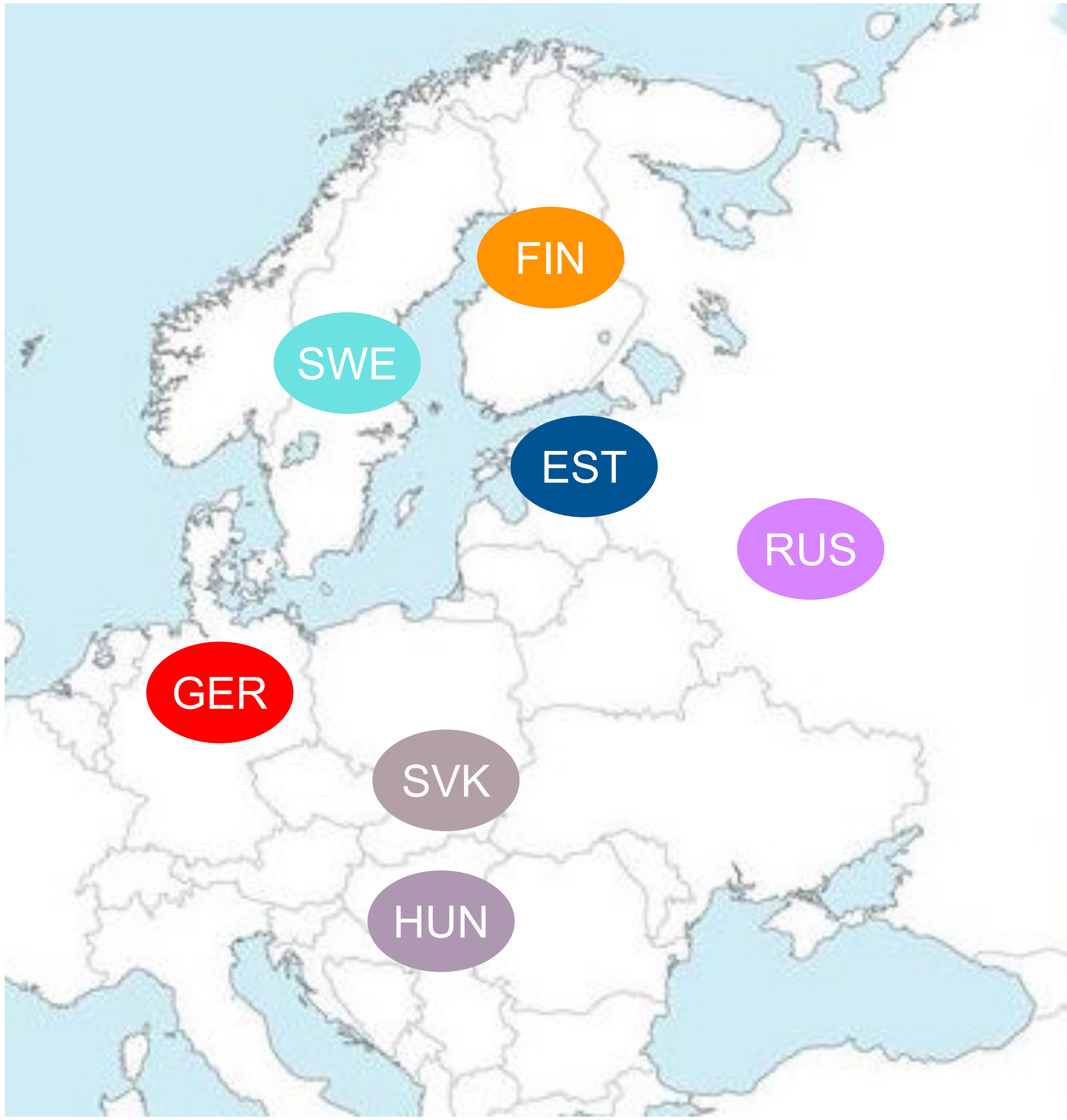




8. Plot something summarizing the confusion matrix

Languages

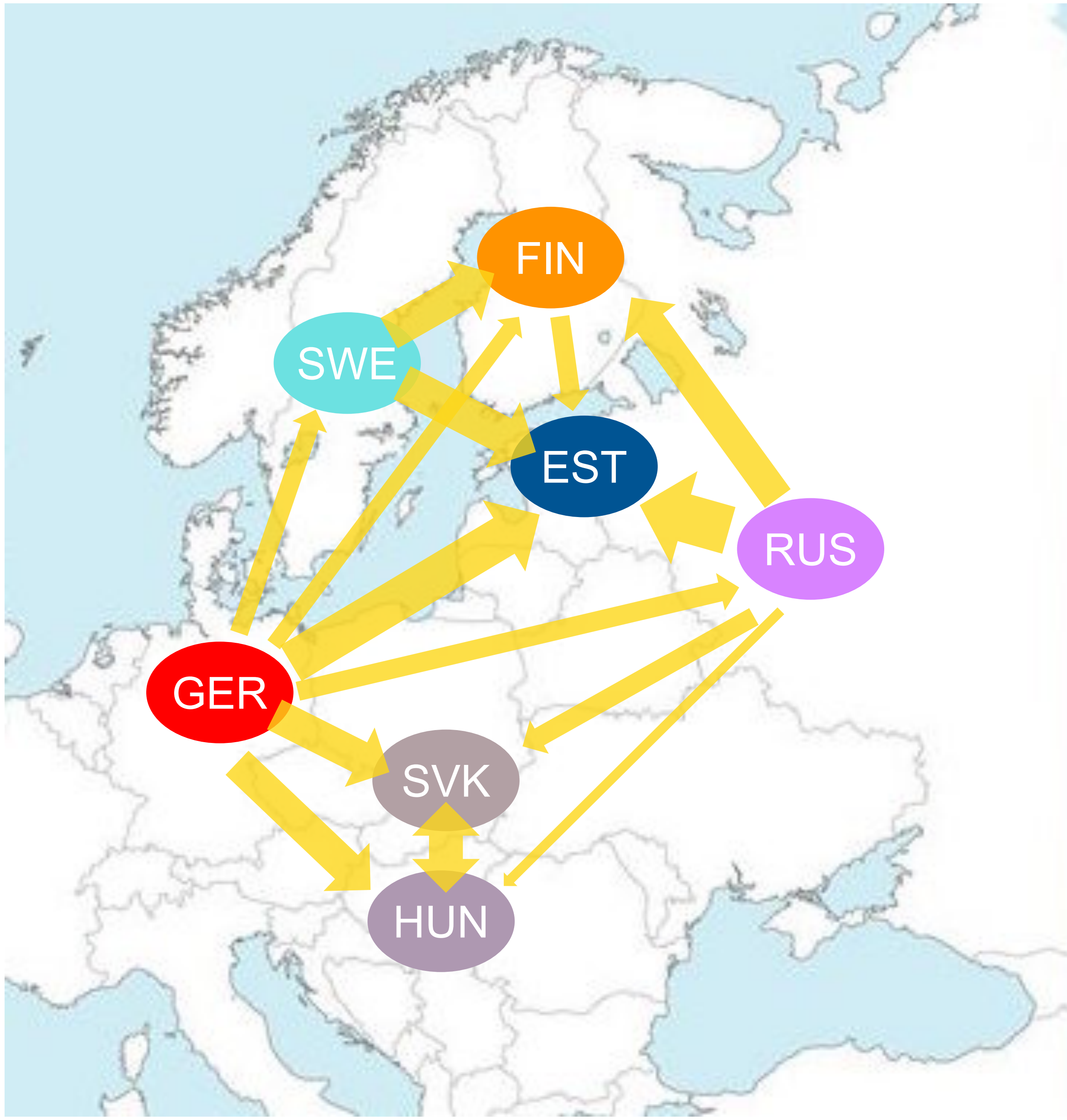
- Seven languages spoken (primarily) in Europe
- 4 Indo-European ones:
  - 2 Slavic (Russian and Slovak)
  - 2 Germanic (German and Swedish)
- 3 Finno-Ugric
  - 2 Finnic (Finnish and Estonian)
  - 1 Ugric (Hungarian)
- Rich and complex mutual contact history



From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

Languages

- Seven languages spoken (primarily) in Europe
- 4 Indo-European ones:
  - 2 Slavic (Russian and Slovak)
  - 2 Germanic (German and Swedish)
- 3 Finno-Ugric
  - 2 Finnic (Finnish and Estonian)
  - 1 Ugric (Hungarian)
- Rich and complex mutual contact history



From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

Languages

Language	Lexical stress	Quantity	Rhythm class	Tone	
Swedish	contrastive	C(2) V(2)	stress-timed	yes	…
German	contrastive	V(2)	stress-timed	no	…
Russian	contrastive	no	stress-timed	no	…
Slovak	word-initial	V(2)	syllable-timed	no	…
Hungarian	word-initial	C(2) V(2)	mora-timed(?)	no	…
Estonian	word-initial	C(3) V(3)	foot-timed(?)	no (?)	…
Finnish	word-initial	C(2) V(2)	mora-timed(?)	no (?)	…

From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

Corpus

- A short story (The North Wind and the Sun), apart from Russian
- Relatively few speakers

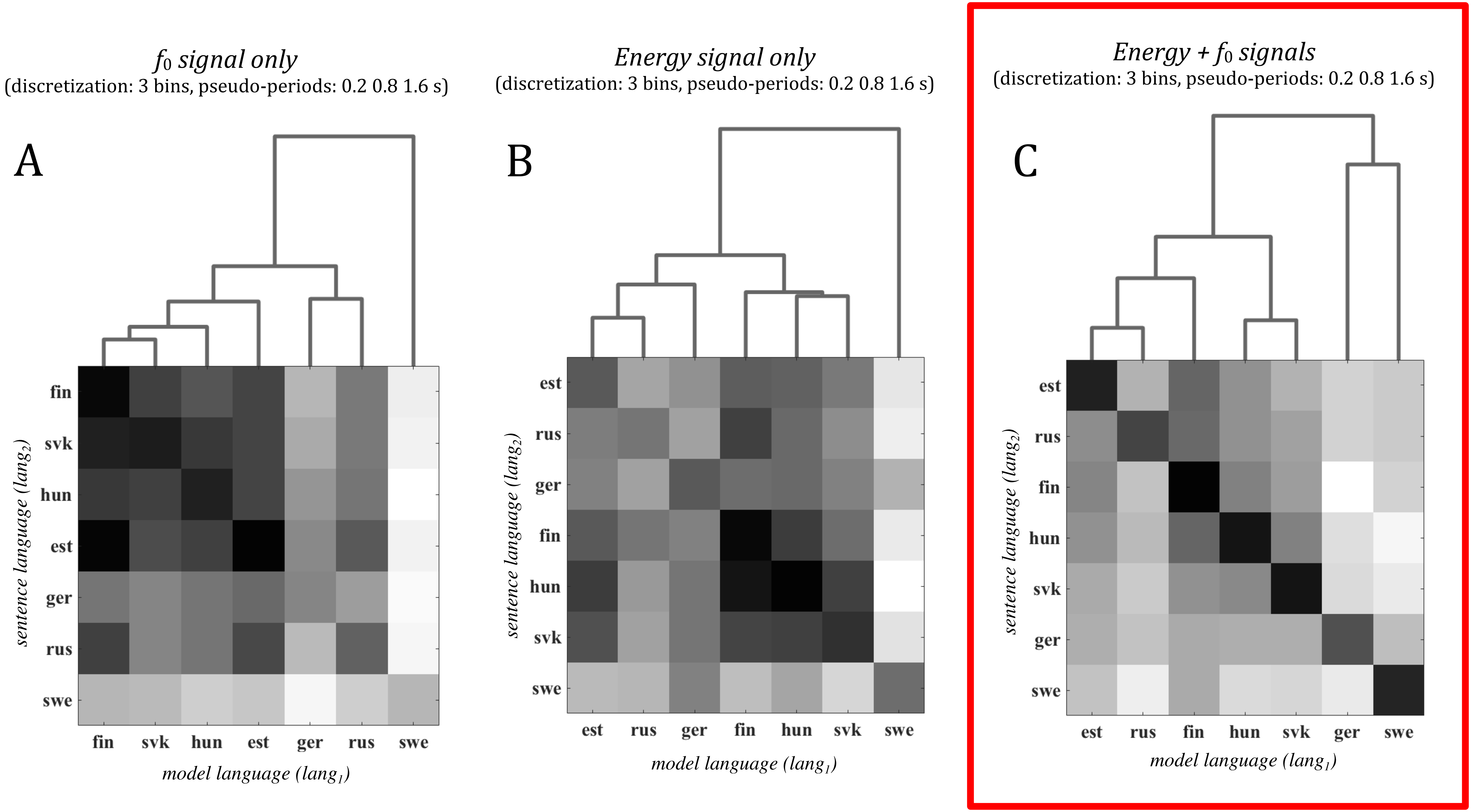
» very small data set for machine learning

Language	Speakers (female)	Sentences	Duration (s)
Swedish	4 (2)	4 x 5	138
German	9 (4)	9 x 5	349
Russian	5 (5)	5 x 10	178
Slovak	6 (3)	6 x 7	176
Hungarian	6 (3)	6 x 7	213
Estonian	6 (3)	6 x 8	207
Finnish	7 (3)	7 x 6	226

less than half an hour

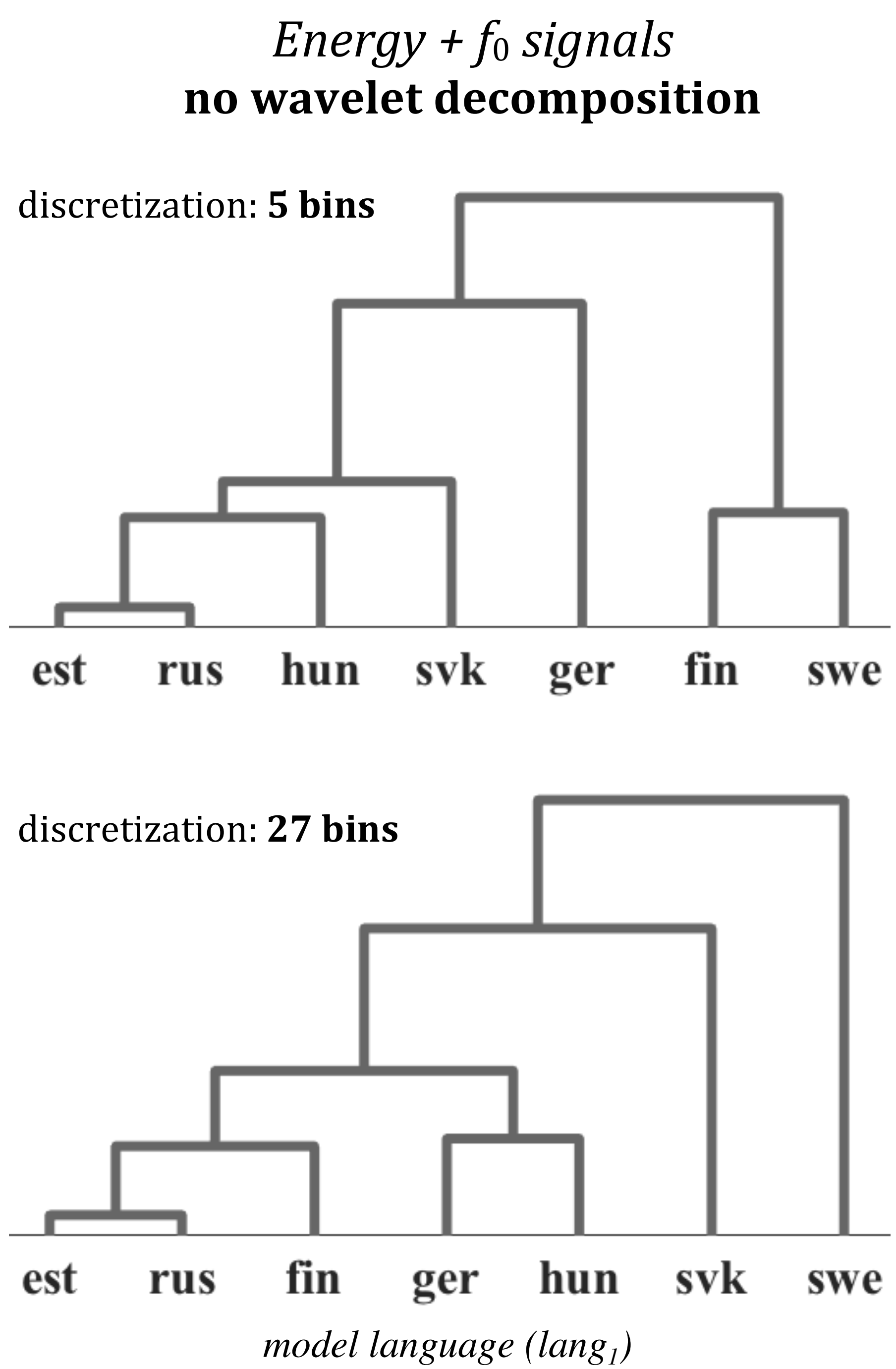
From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

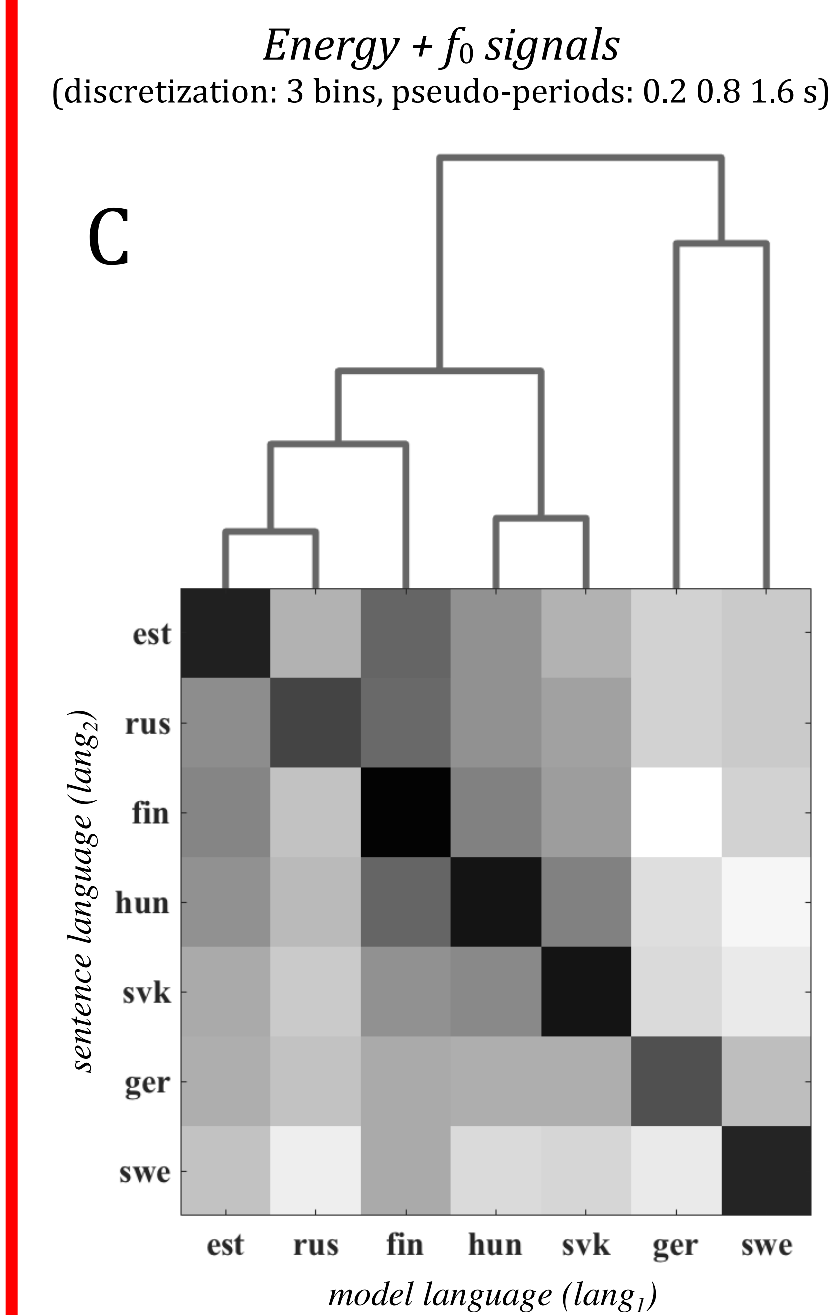
Results: CWT decomposition



From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

Results: No CWT decomposition

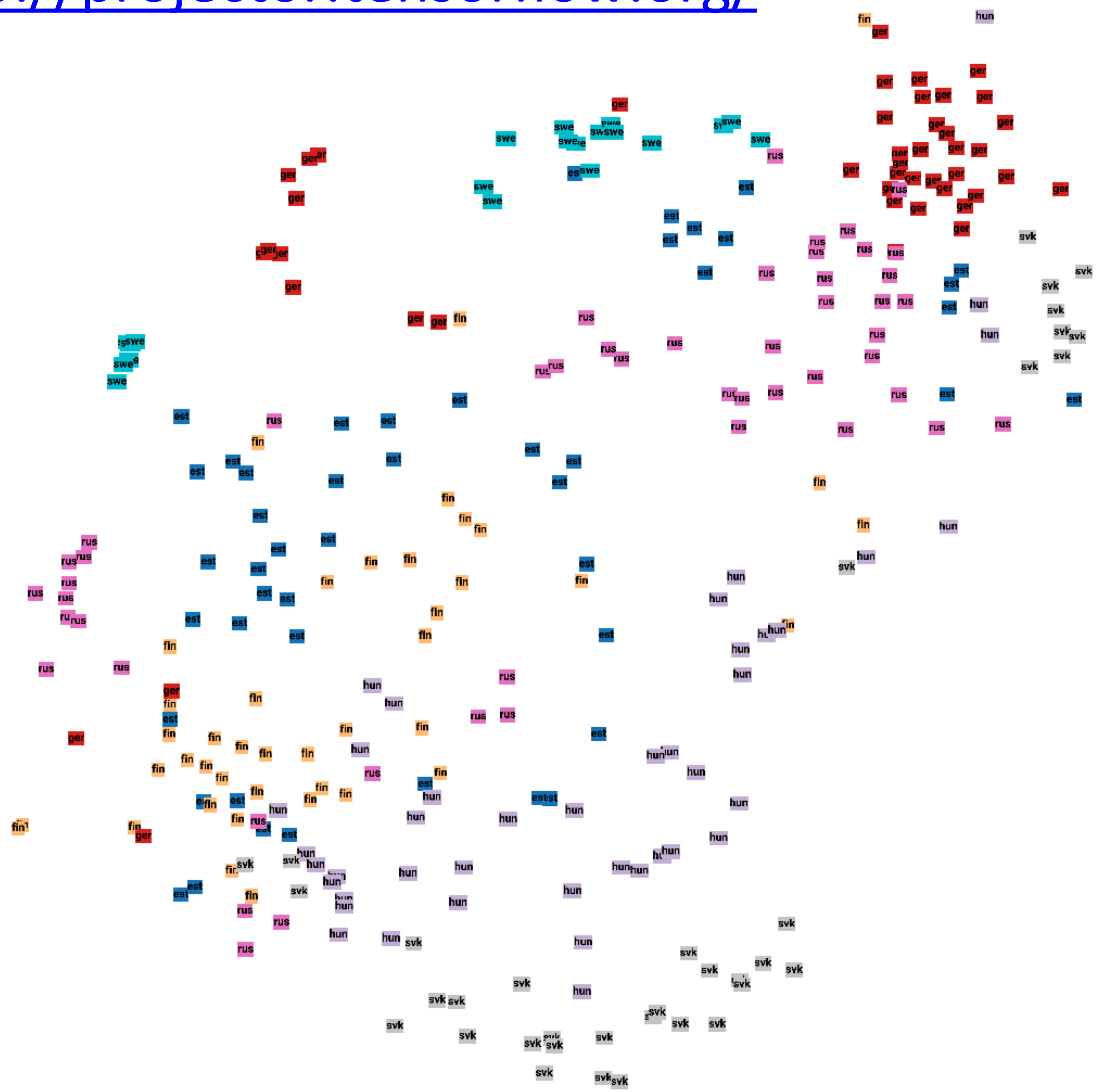


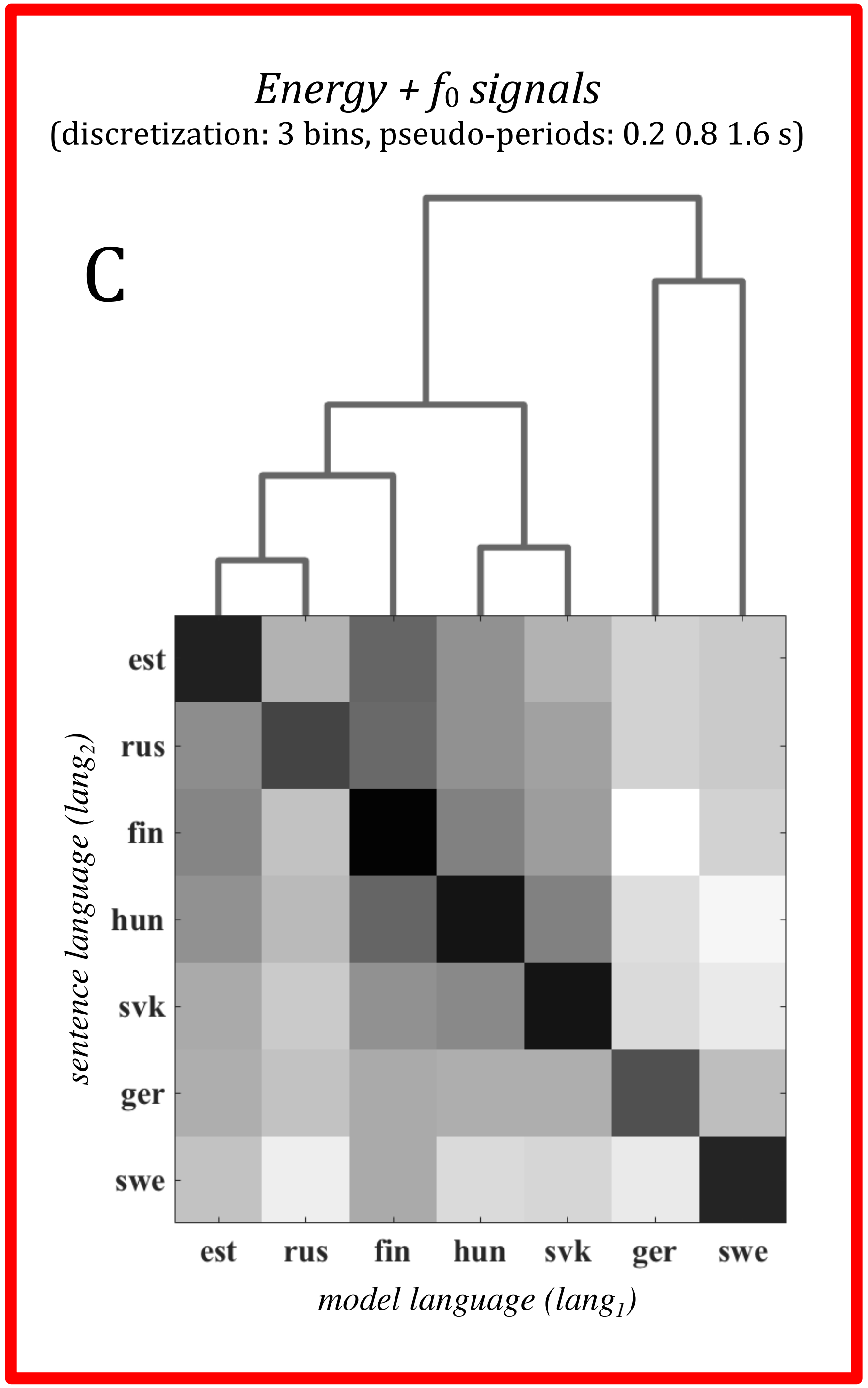


From: Šimko, Suni, Hiovain, Vainio (2017, Interspeech)

Another way to look at it

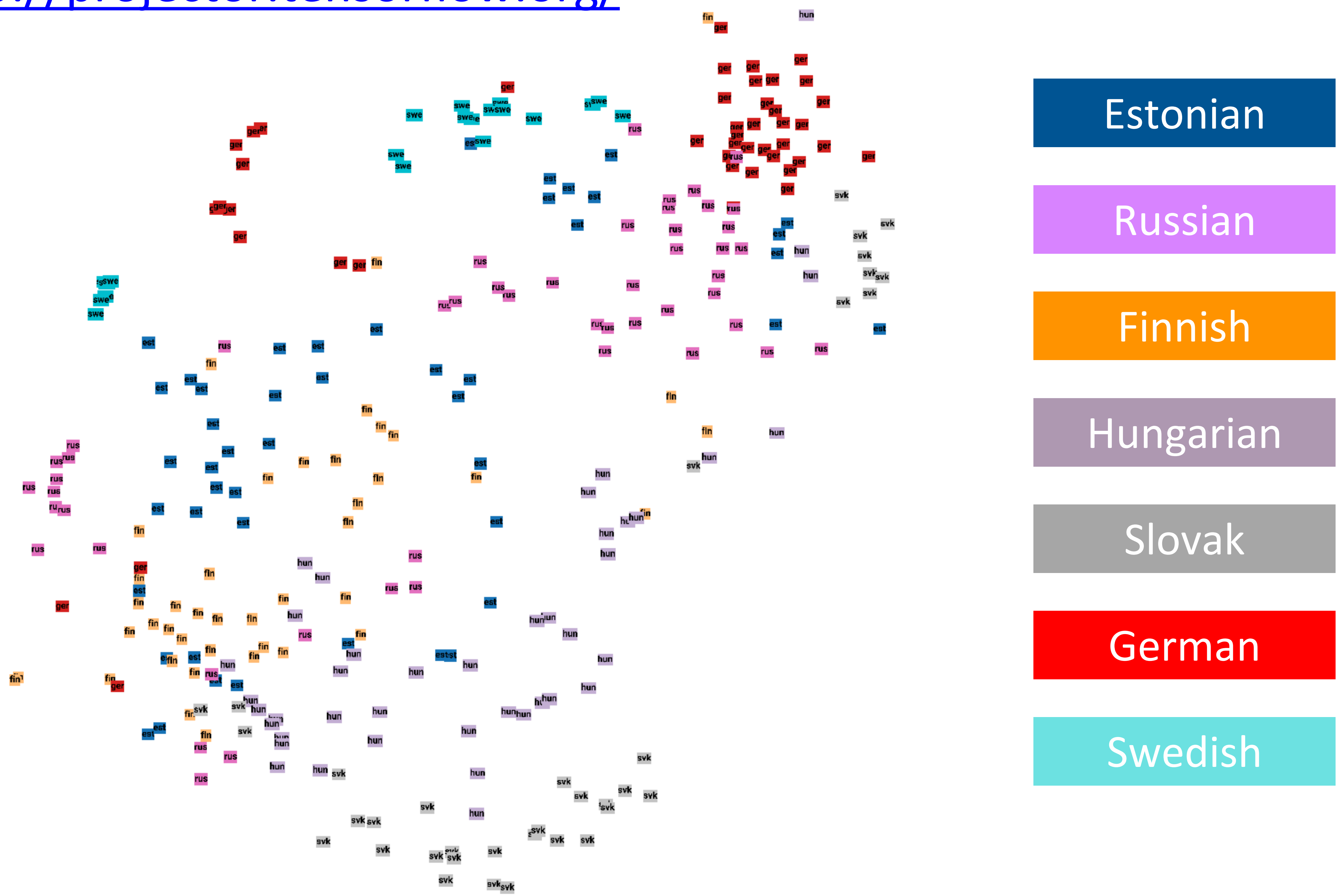
http://projector.tensorflow.org/





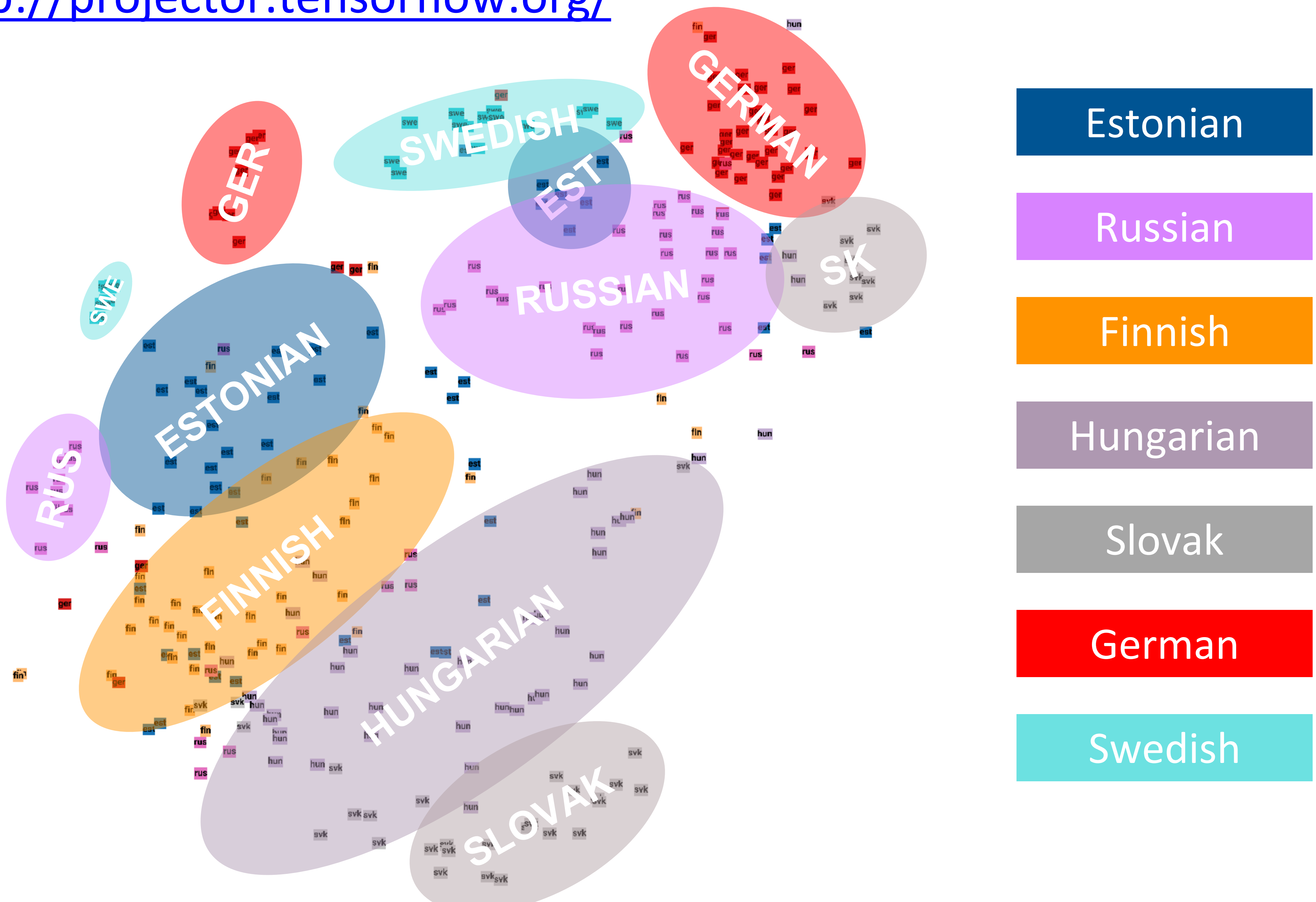
Another way to look at it

http://projector.tensorflow.org/



Another way to look at it

http://projector.tensorflow.org/



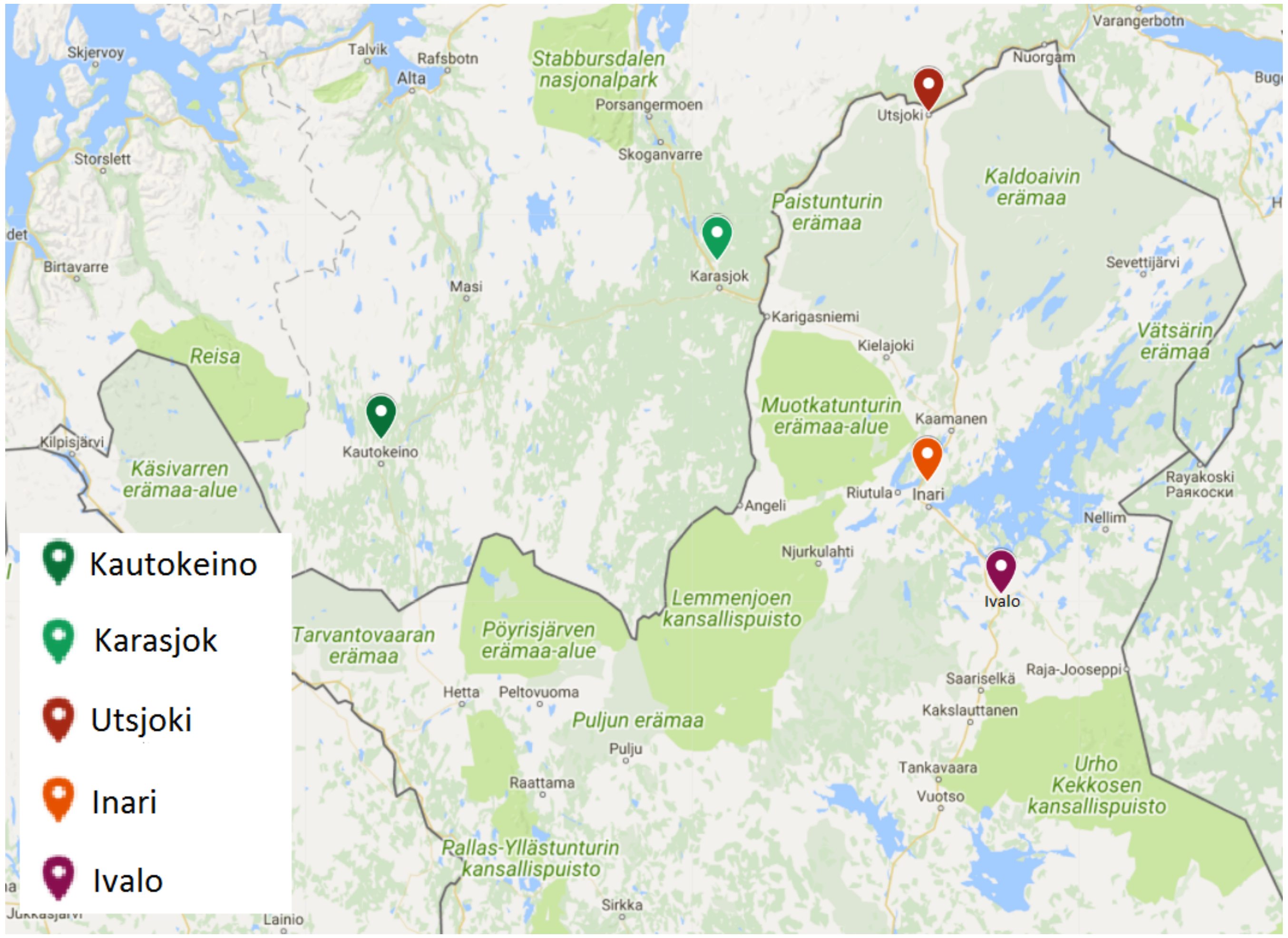
Another, slightly bigger corpus

- **North Sámi**

NS varieties	Spkrs (female)	Minutes
Kautokeino (**skt**)	5 (2)	75:09
Karasjok (**skr**)	6 (5)	43:02
Ivalo (**siv**)	6 (5)	43:29
Utsjoki (**sut**)	6 (2)	86:30
Inari (**sin**)	4 (3)	43:54

Majority lgs	Spkrs (female)	Minutes
Finnish (**fin**)	1 (0)	11:47
Norwegian (**nno**)	1 (0)	13:32

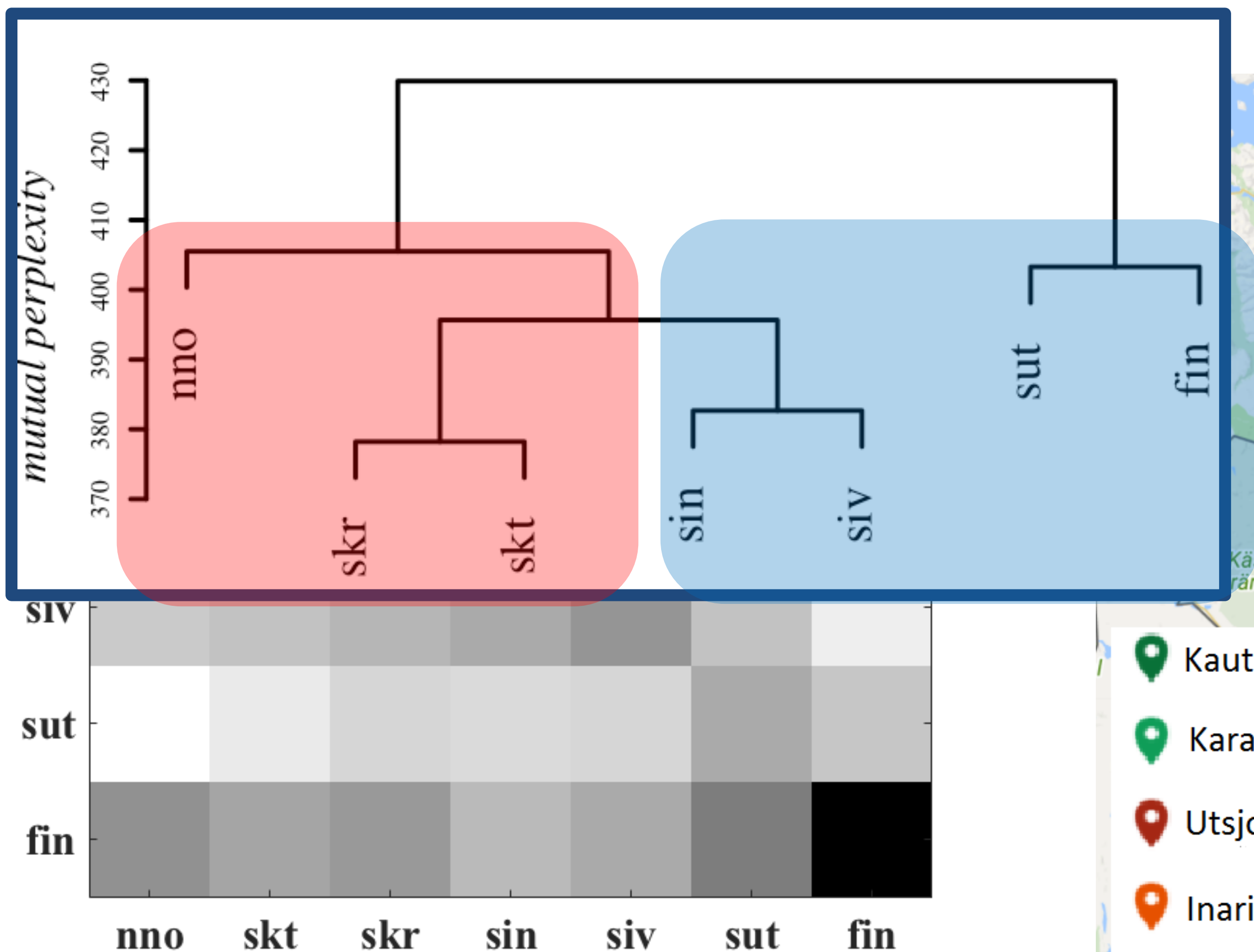
a bit over 5 hours of speech

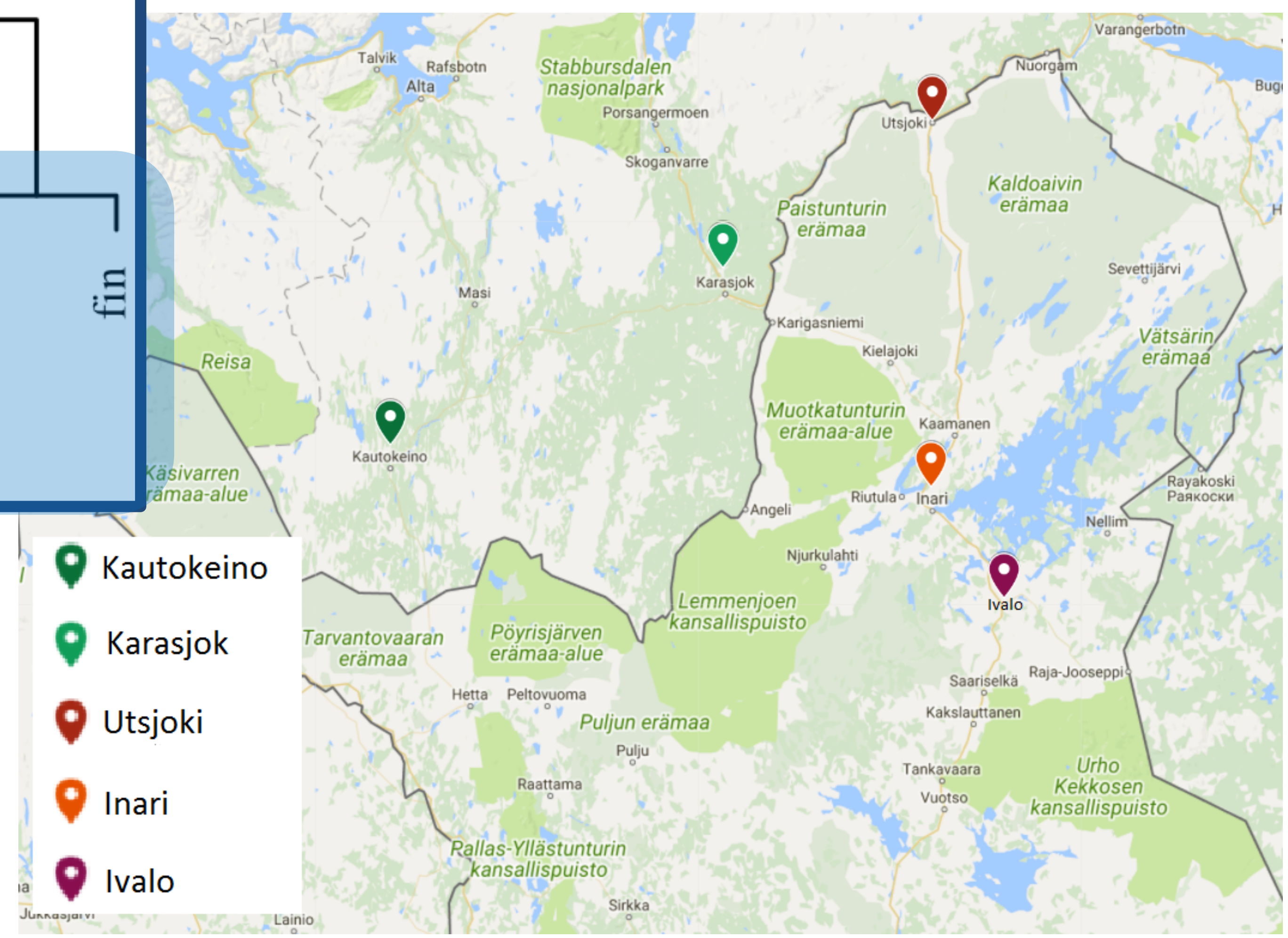


Submitted: Hiovain, Suni, Vainio, Šimko (2018, Speech Prosody)

Another, slightly bigger corpus

- **North Sámi**





Yet another, even bigger corpus

- SWEDIA 2000 (Bruce, Elert, Engstrand, Eriksson and Wretling, 1999)
- **in Swedish**
- individual words from 104 locations from Sweden and Finland, different dialects

(lot of words) * (lot of speakers) = = over 250,000 renditions

= about 2 days of words!

(1.2 million files processed)

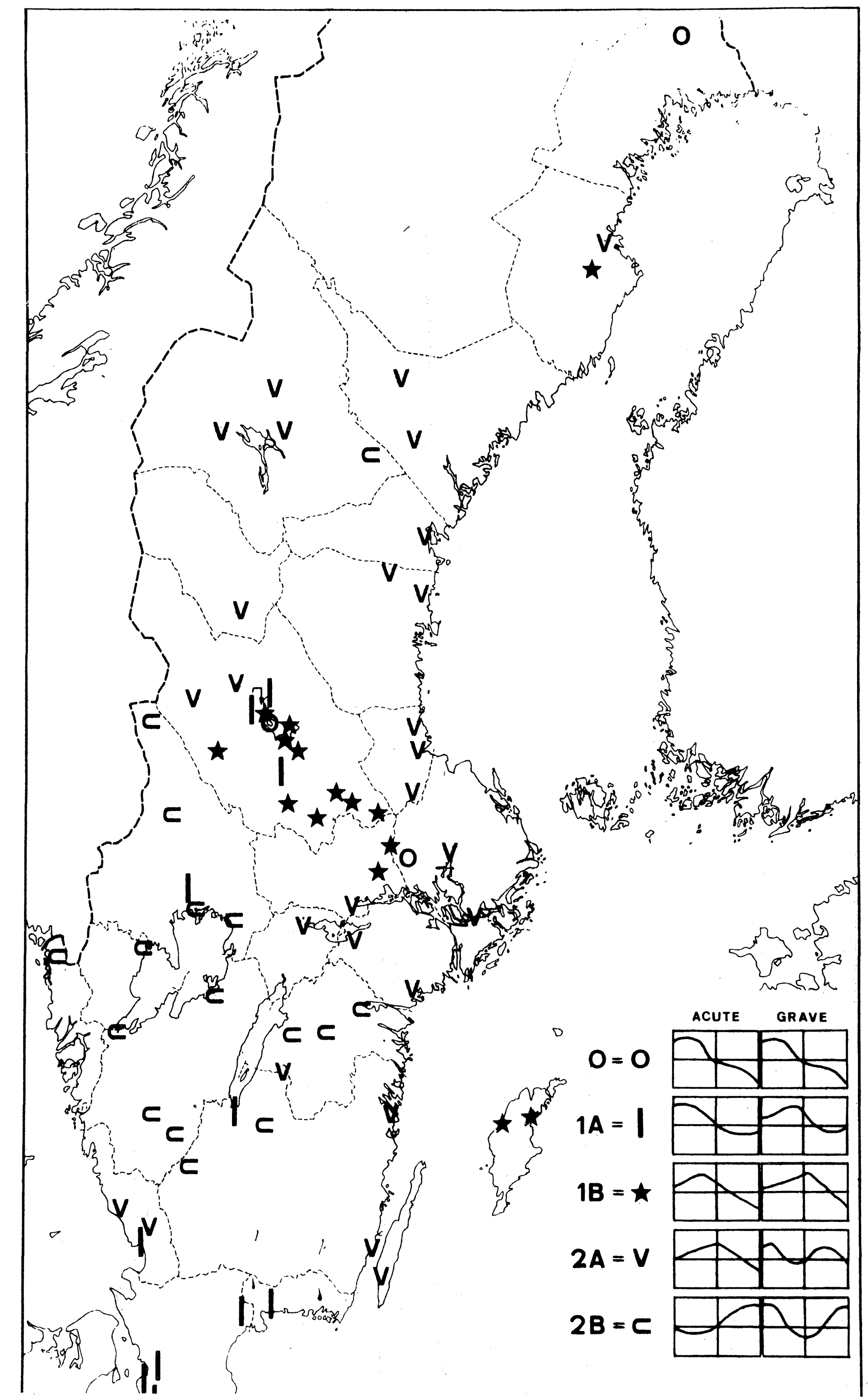
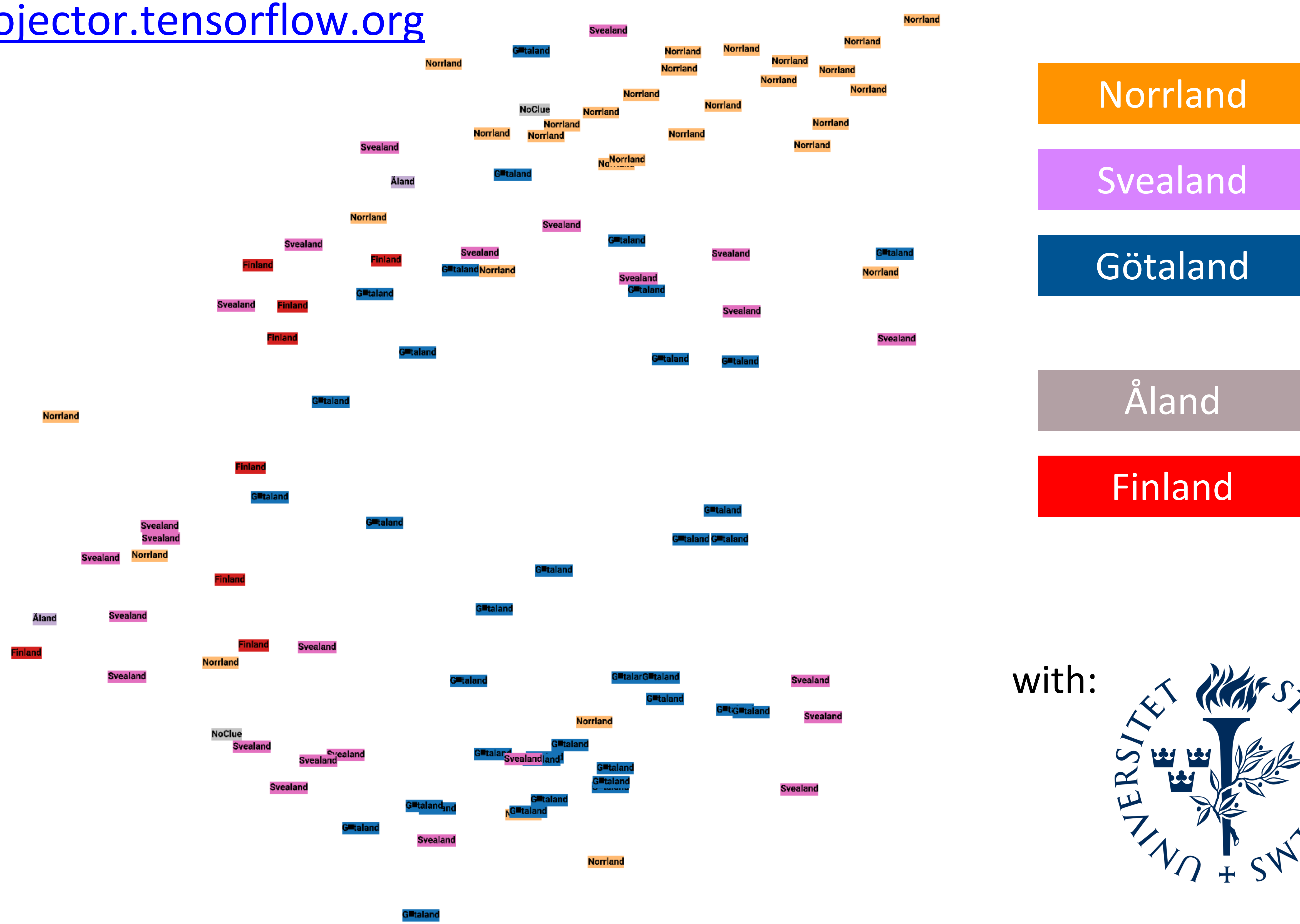


Figure 1. Geographical distribution of the accent types (from Gårding & Lindblad 1973).

SWEDIA 2000 dialects

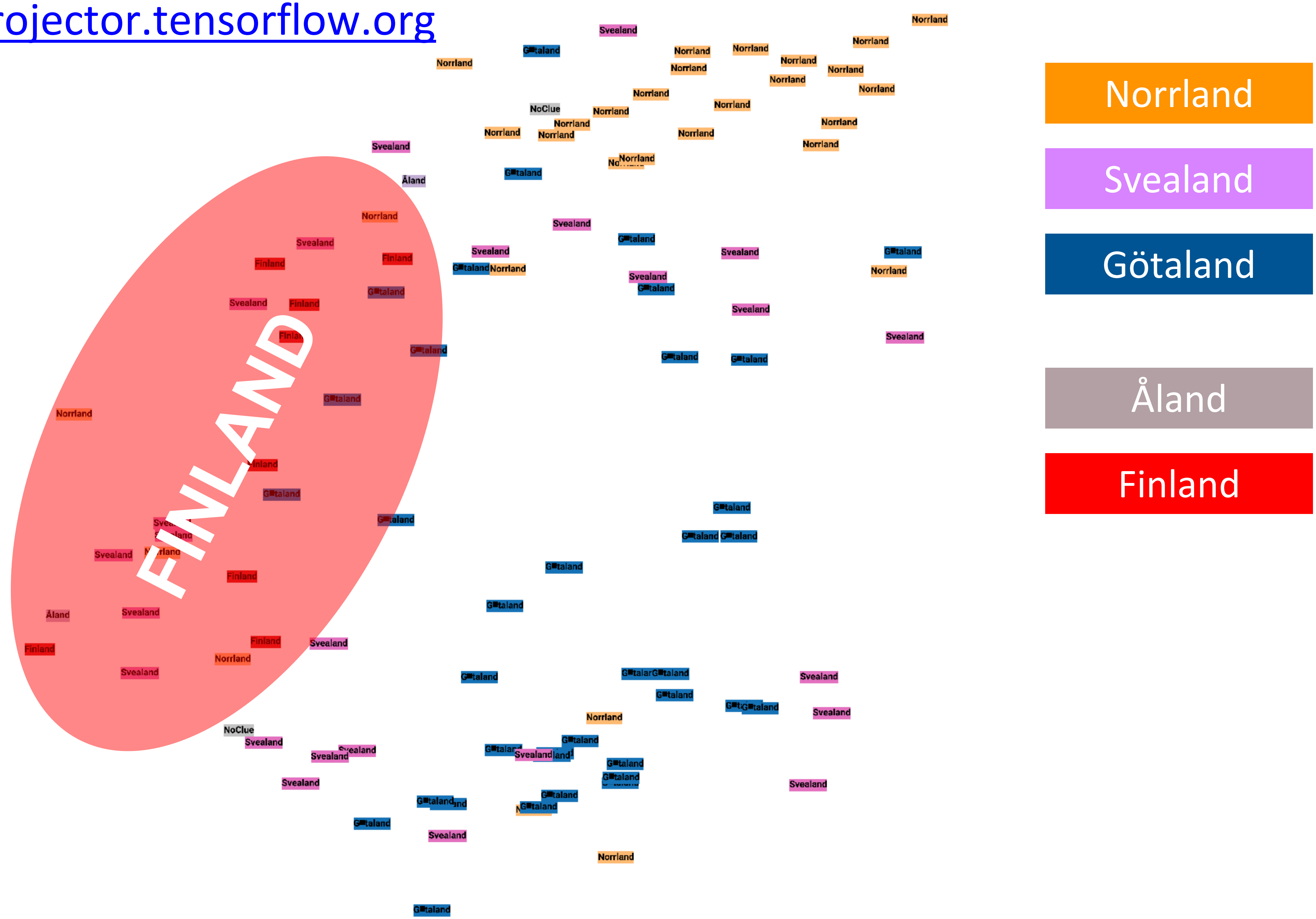
projector.tensorflow.org



with:

SWEDIA 2000 dialects

projector.tensorflow.org



Discussion

- very simple language modeling (unigrams)
  - with bigger corpus, we will (and do) try more complex modeling, e.g., deep nets

- are our results “right”?
  - lack of the Ground Truth
  - instead, we need to compare the known characteristics of the languages and use common sense

Discussion

- works for both small and big corpora
- the results seem to be meaningful:
  - the language grouping largely reflects language family relationships (**fin-est; swe-ger**), and contact history (**svk-hun**)
  - Swedish dialects “sort out” in geographically meaningful(ish) way
  - North Sámi data also seem to make sense
- wavelet decomposition helps
  - statistical evaluation of f_0 and energy envelope movement distribution patterns on multiple hierarchical levels **in parallel** (inter-dependencies) seem to capture relationships better than simple raw contours
- combined signals (energy+f_0) give “more plausible” results than each signal separately (cf. Cummins etal., 1999)

*Antti Suni, Katri Hiovain, Martti Vainio, Atte Hinka,
Mark Granroth-Wilding, Hannu Toivonen*

kiitos ďakujeme aitäh thanks