Digital Language Typology

Mining from the Surface to the Core

Juraj Šimko

Institute of Behavioural Sciences

University of Helsinki

—

BAULT 2016

University of Helsinki

01.12.2016

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Typology

- **Grouping** of languages according to their characteristics
- **Explaining** distributions, language contact
- **Multi-dimensional** space of similarities / differences / influence of contact: syntax, morphology, phonotactics, prosody, ...

Finnish --- Hungarian

Swedish --- Finnish Swedish --- Finnish

Hungarian --- Slovak

Digital (Language Typology)

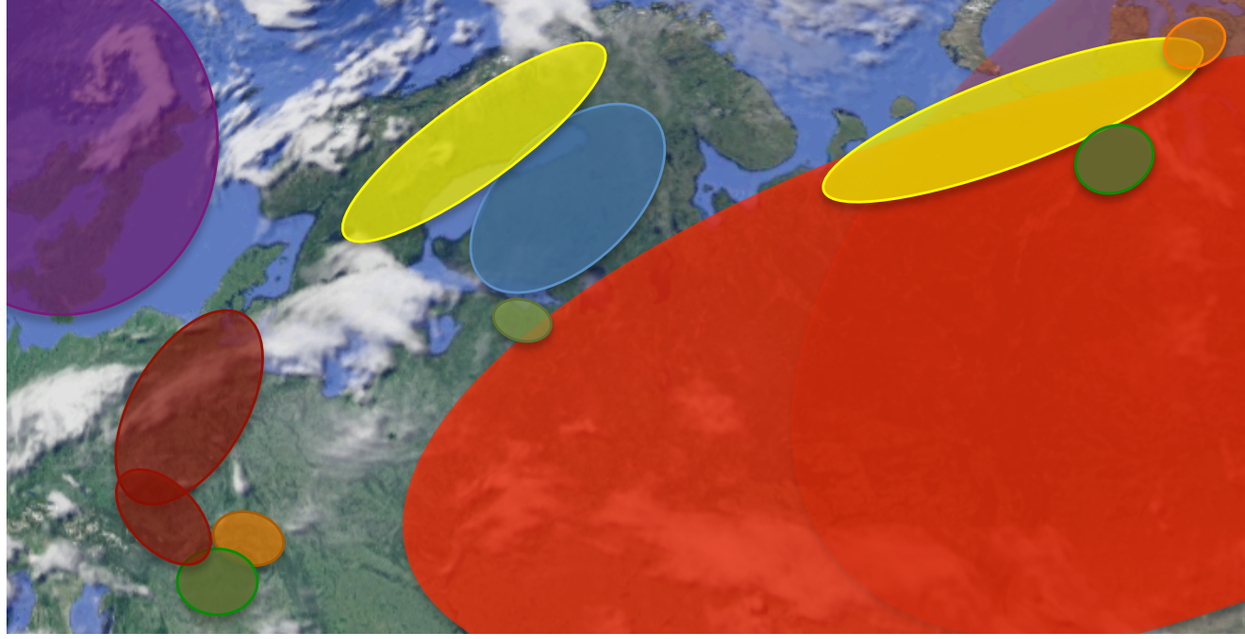
- using language/speech technology tools
- shallow, but non-trivial analysis

and

(Digital Language) Typology

- digital humanities project
- big, digital, language and speech data
- smaller data sets for sanity checks

Languages

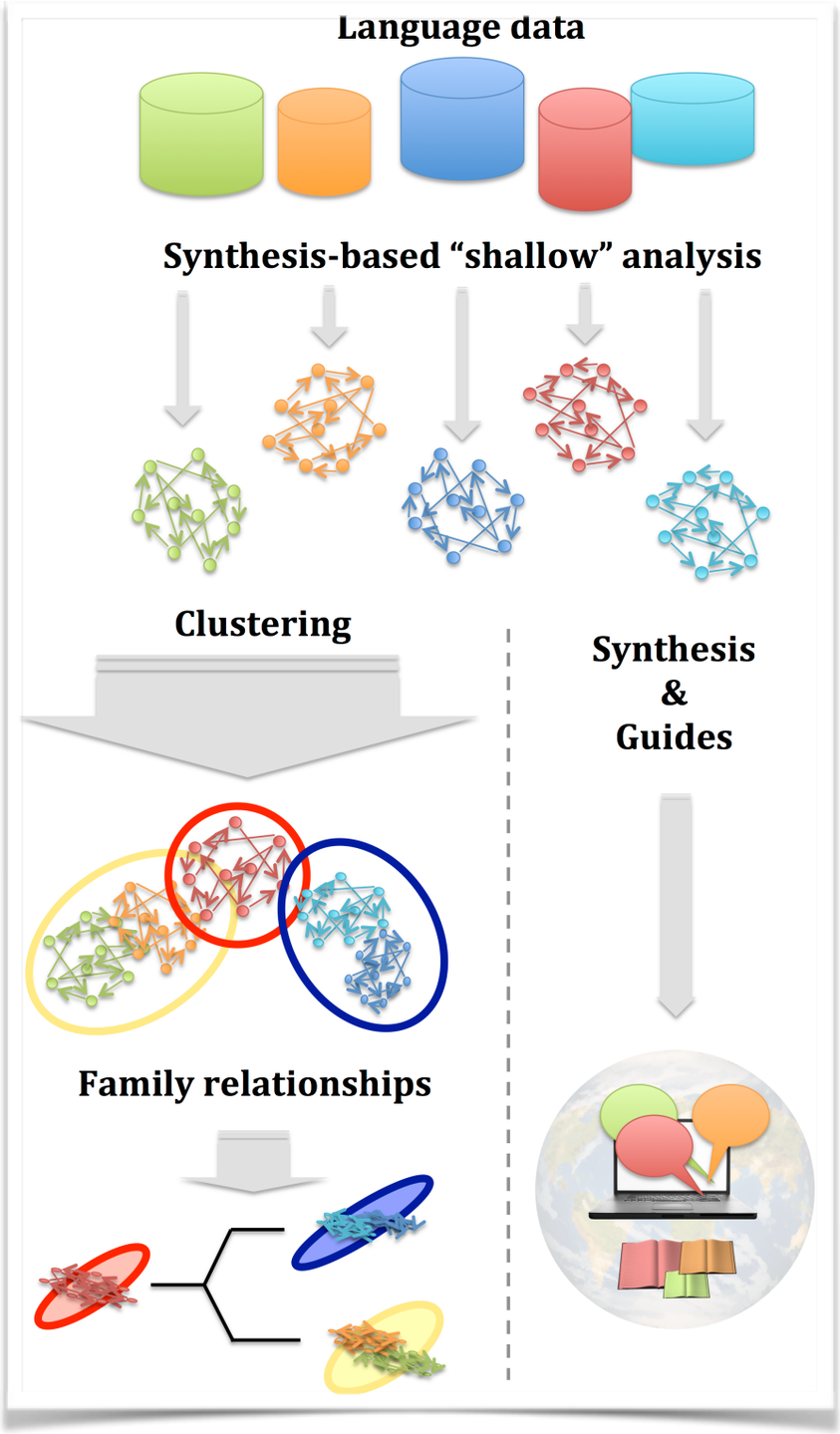


- **Fenno-Urgic:** Finnish, Estonian, Hungarian, Tundra and Forest Nenets, Nganasan and *North Saami*
- **Slavic:** Russian, Slovak, *Czech*
- **Gemanic:** Swedish, German, English, *Norwegian, Danish*
- **Other:** *Latvian*, *Lithuanian*,...

Consortium

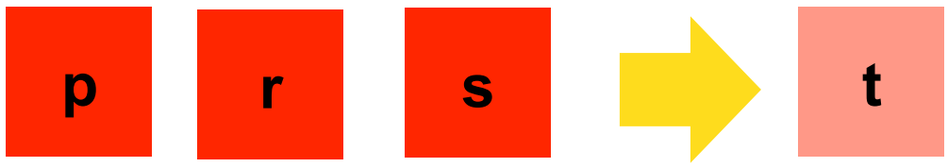
- **UH Phonetics,** PI Martti Vainio:
  - Juraj Šimko
  - Antti Suni
  - Katri Hiovain
- **UH Comp. Science,** PI Hannu Toivonen
  - Mark Granroth-Wilding
  - Atte Hinkka
- **UTA Info. Sciences,** PI Markku Turunen
  - Larisa Leisiö

Project Outline



Language n-grams and perplexity

$p_{FIN}(t|(t,a,m,...))$

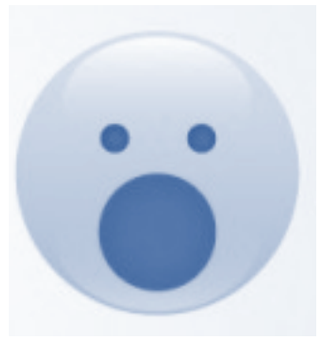


$p_{SVK}(t|(s,r,p,...))$

Language n-grams and perplexity

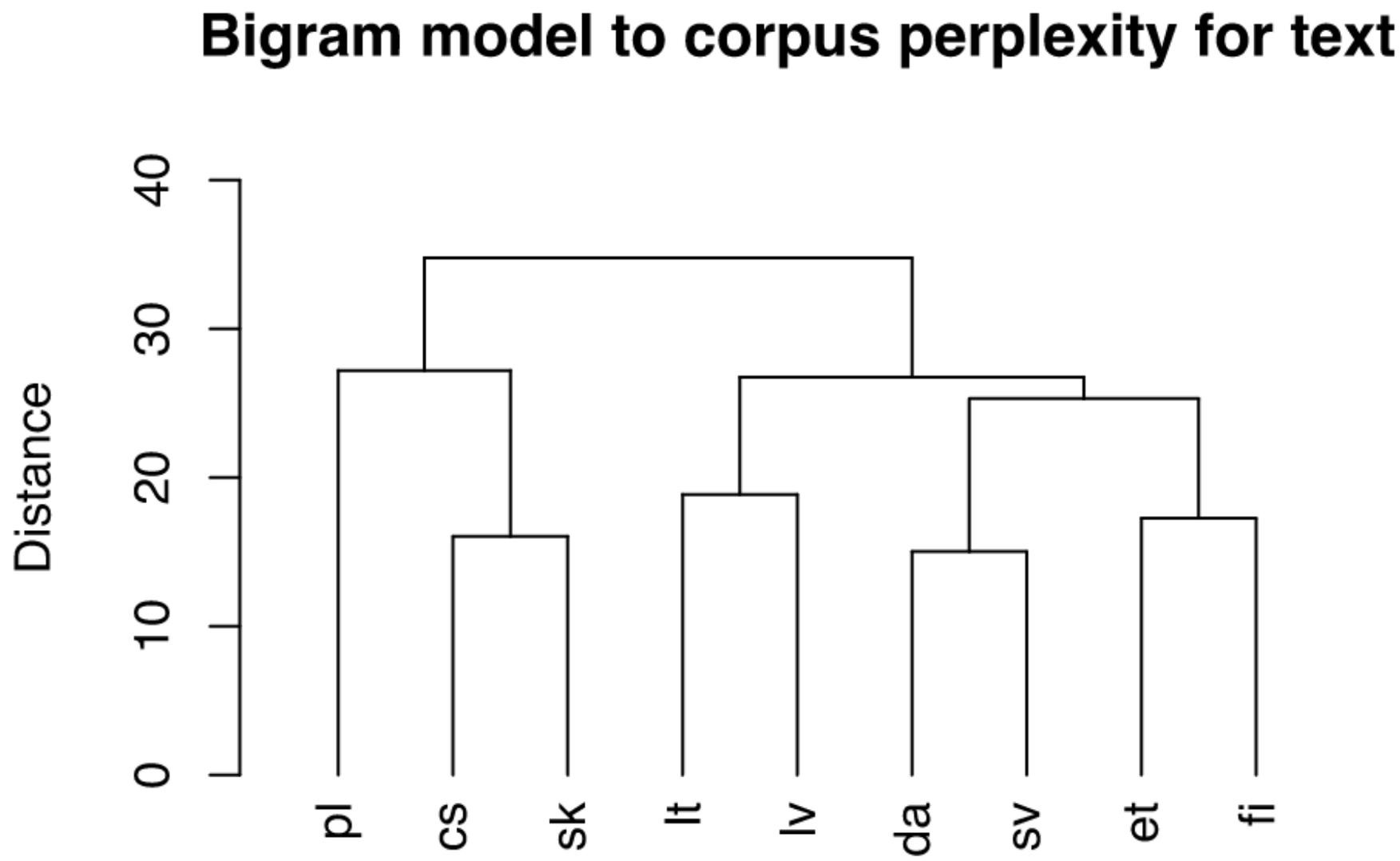
$p_{SVK}(t|(t,a,m,...))$

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



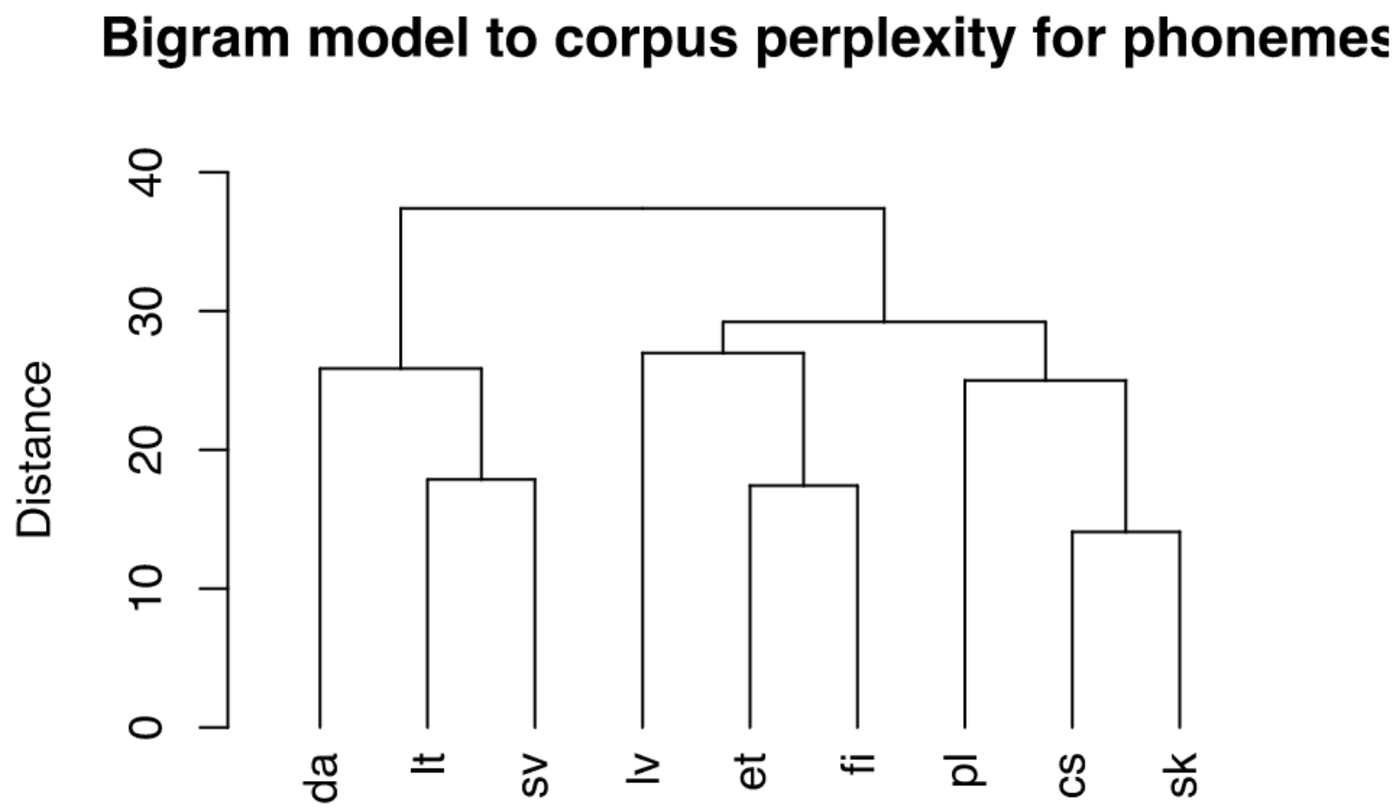
Language n-grams and perplexity

- Using the EU Europarl corpus, standard orthograpy



Language n-grams and perplexity

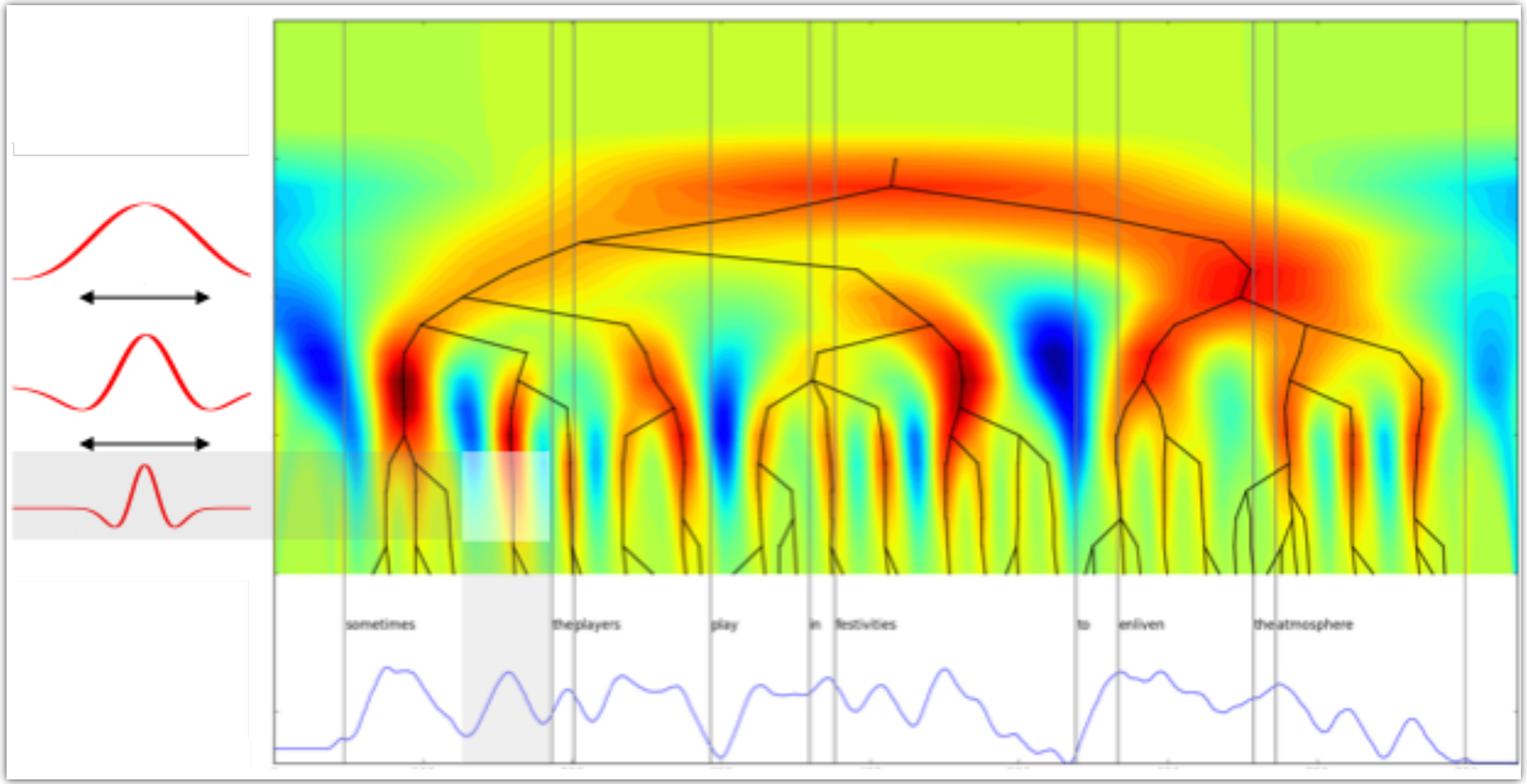
- Same corpus, transcribed using espeak



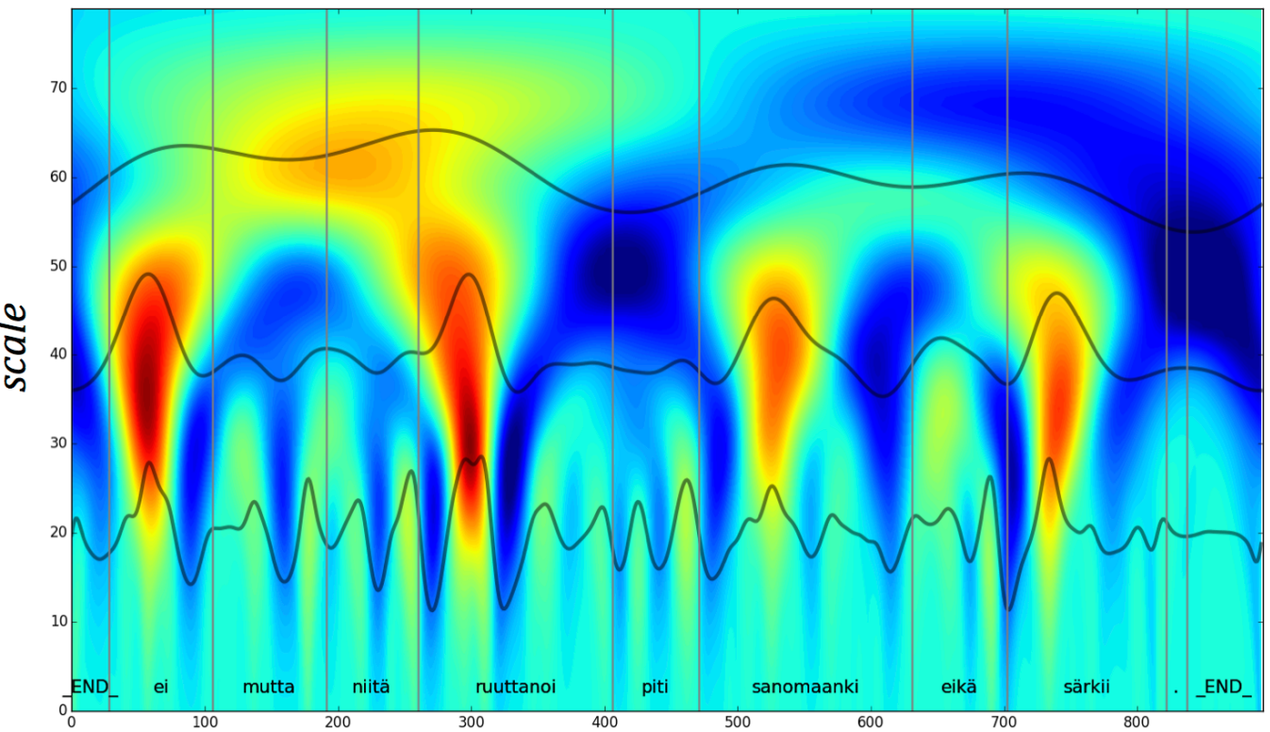
- Not so good, non-matching phoneme sets
- We can see where the models are most perplexed: **sanity checks**

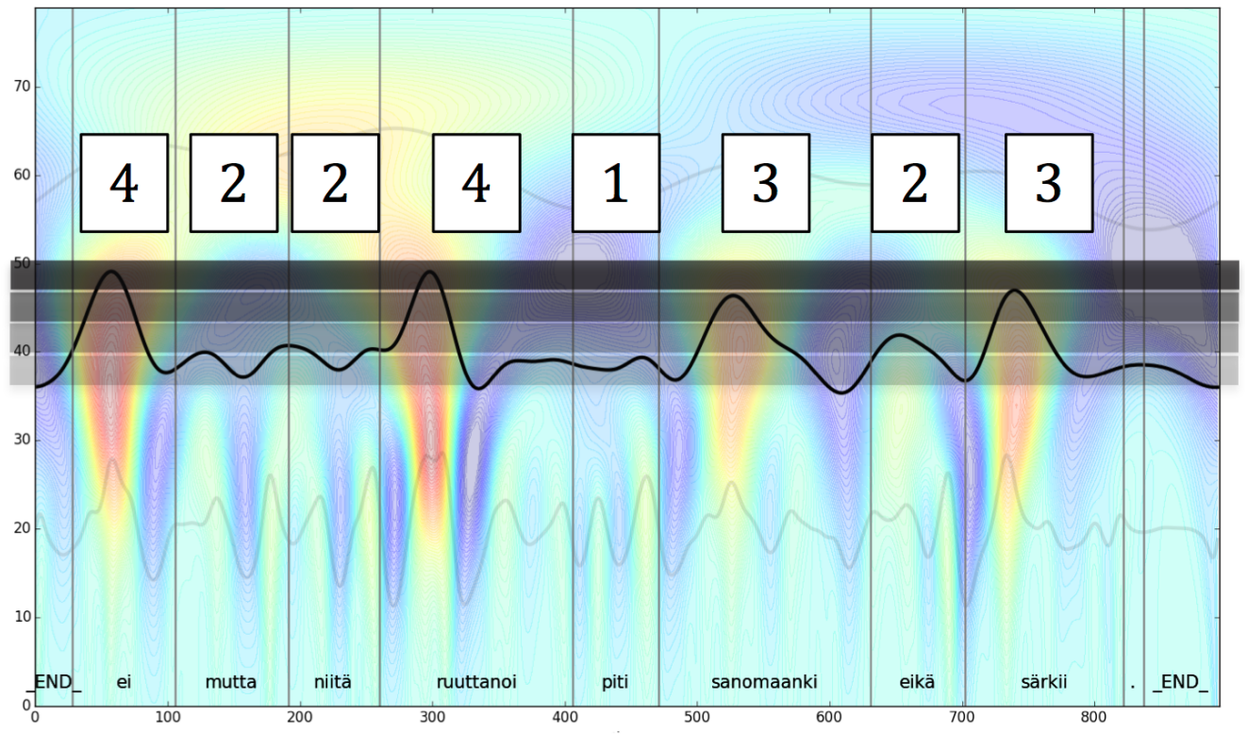
Prosody

- Speech is structured hierarchically (phrases -> (phonological) words -> syllables -> speech sounds -> acoustic events)
- Hierarchical analysis: Continuous Wavelet Transform



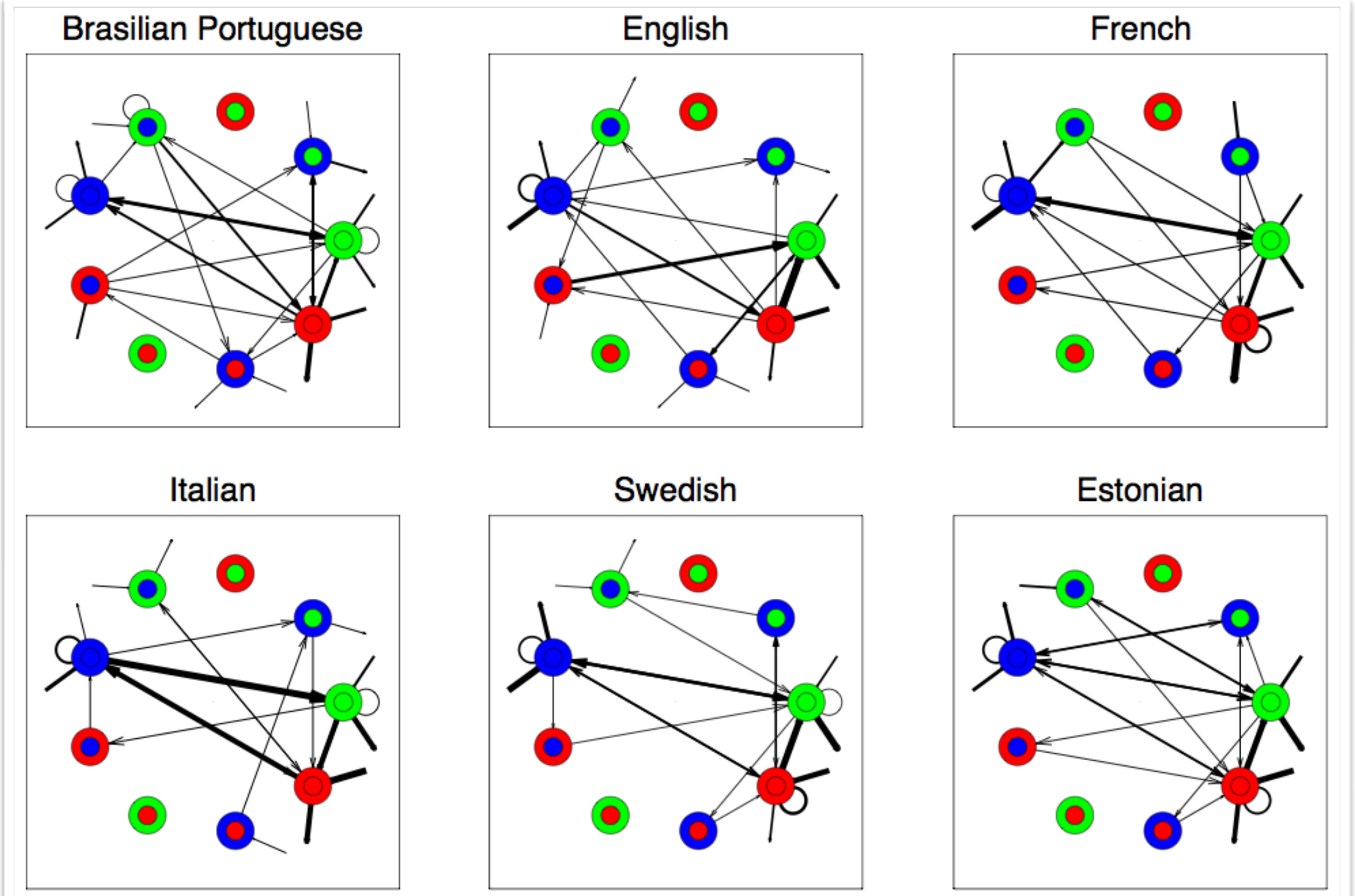
Estimating prominences



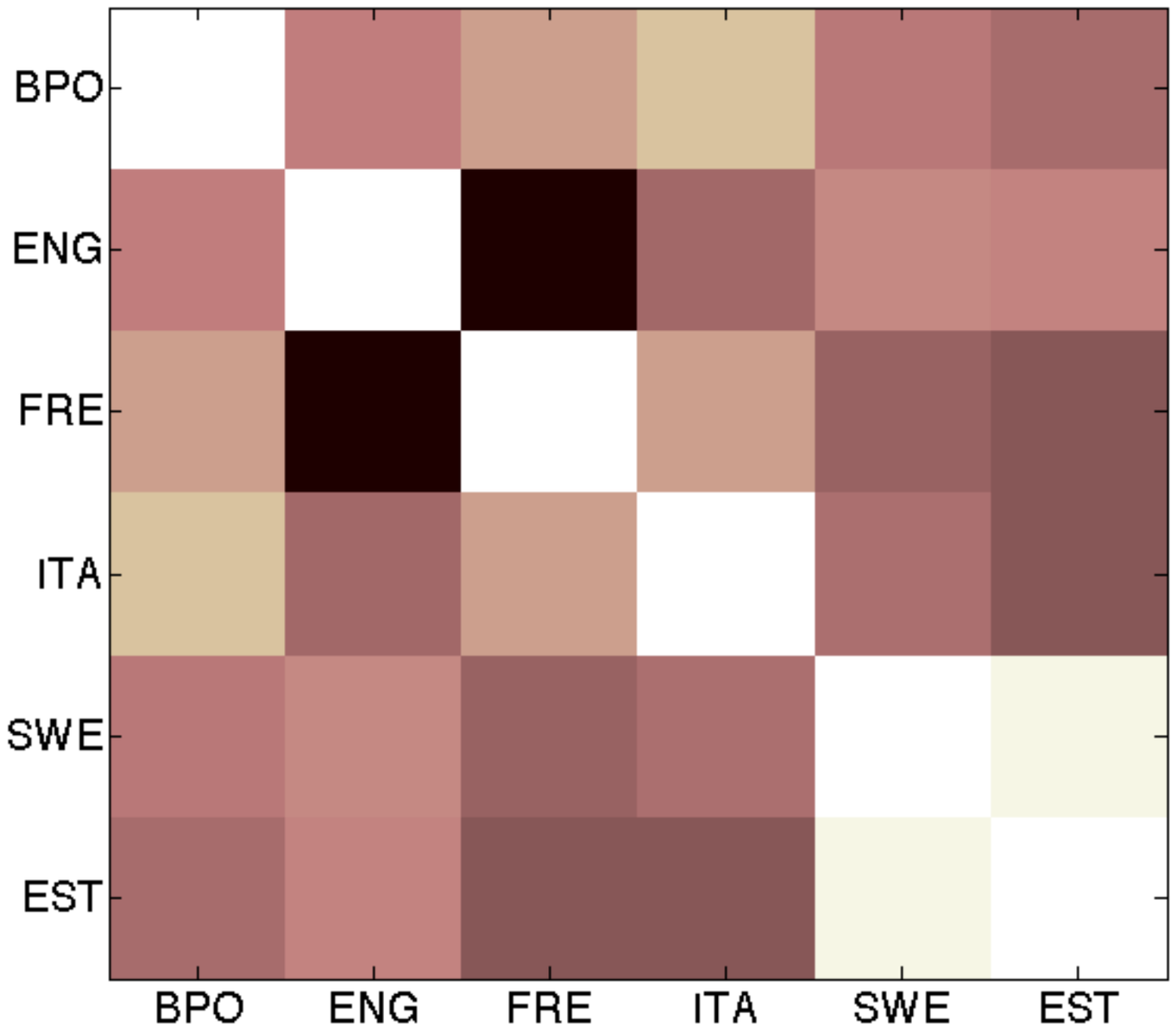


- CWT analysis can be used to estimate prominences and boundaries at several levels
- The prosodic structure can be discretised on several relevant linguistic levels ... *and fed to an n-gram model*

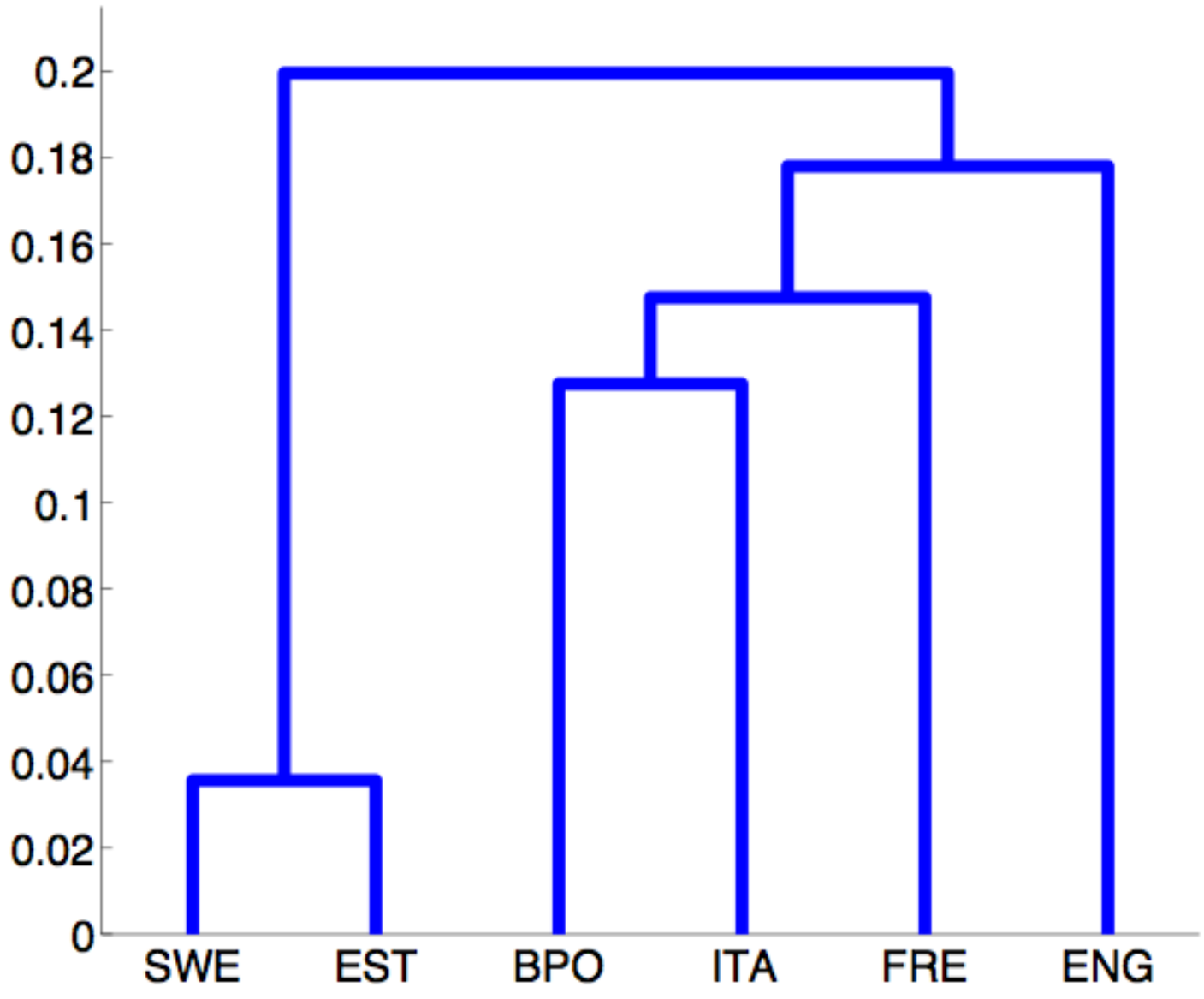
Transition probabilities on a mini corpus



Similarity



Prosodic dendrogram

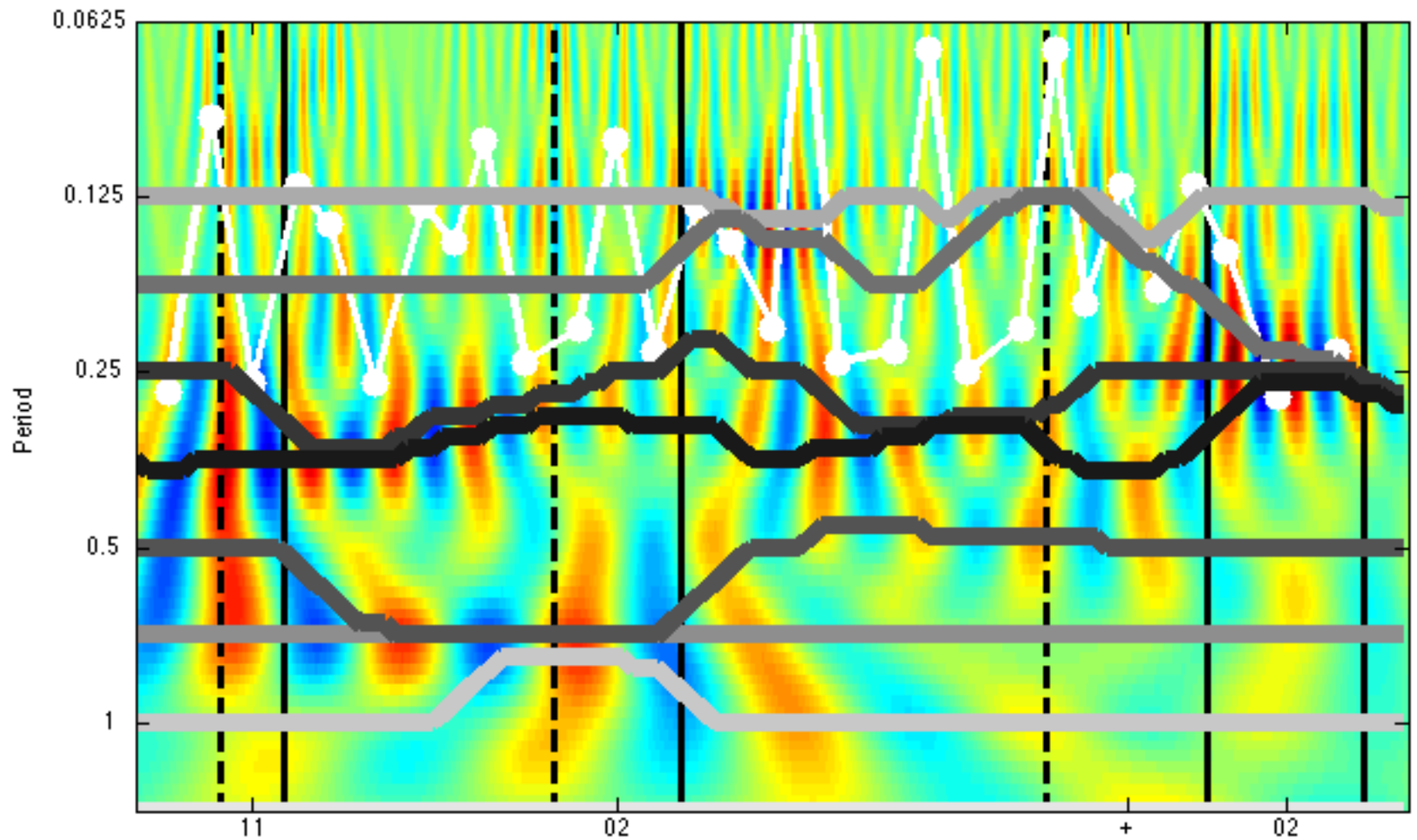


Now

- **Exploring and fine-tuning methods:**

  n-grams -> RNN

  tracking prosodic events



Now

- **Collecting small-ish speech and language corpora**
  annotated "The North Wind and the Sun" in many languages
  corpora from small languages:
  North Saami, Samoyed **PLEASE HELP**

- **Expanding our interests somewhat**
  e.g., dialects: Saami spoken in Finland and Norway

- **Getting to know each other better**

Near future

- **Morphology, syntax**
  Morfessor, SyntaxNet,....

- **Doing the work**
  building models
  comparing them
  clustering
  visualizing

- **Building speech synthesis systems for small under-resourced languages**
  Samoyed, Saami,...

Keep fingers crossed!

kiitos *ďakujeme* *спасибо*

thanks *aitäh*

Keep fingers crossed!

kiitos ďakujeme спасибо

thanks aitäh