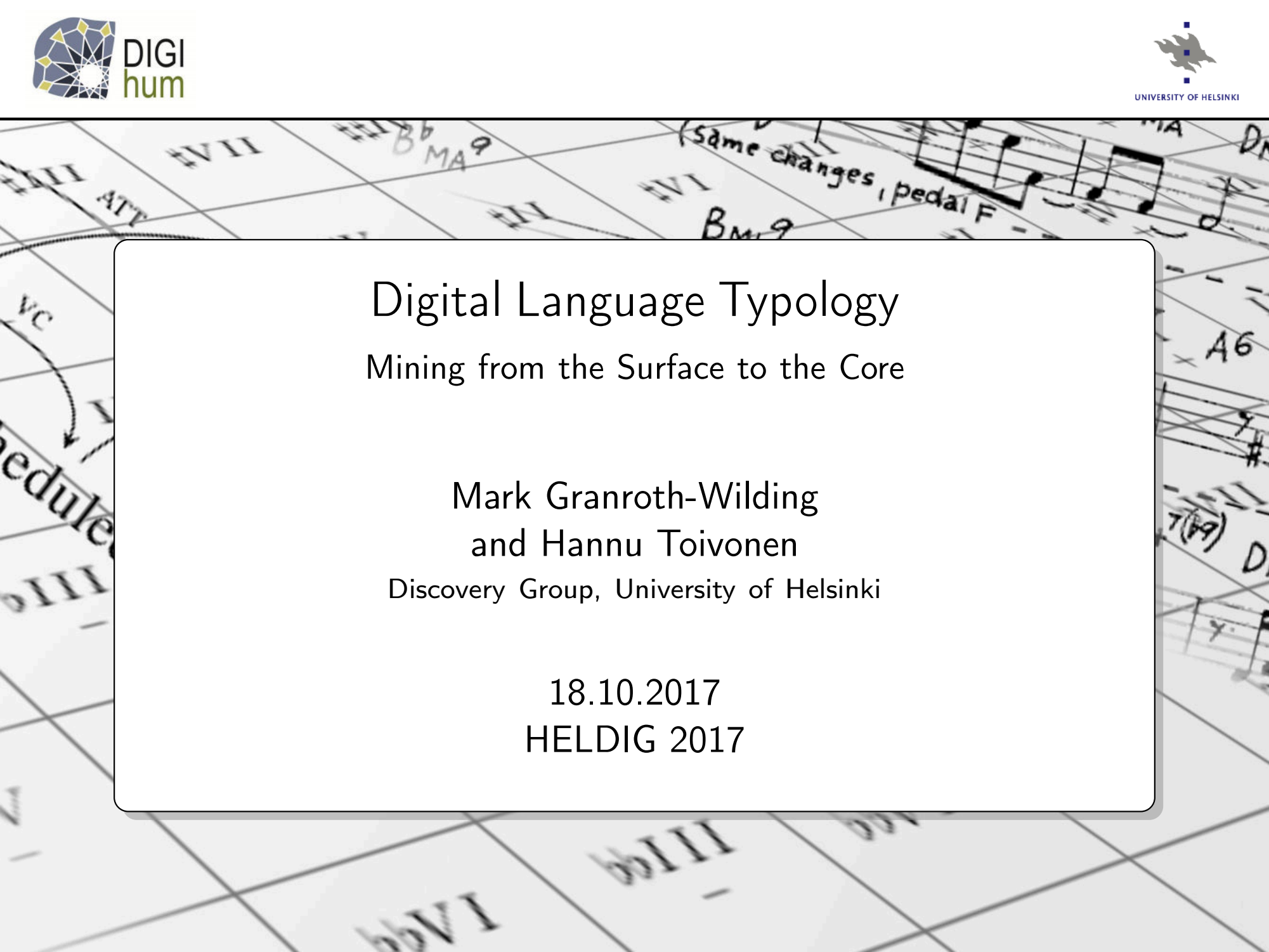


Digital Language Typology

Mining from the Surface to the Core

Mark Granroth-Wilding
and Hannu Toivonen
Discovery Group, University of Helsinki

18.10.2017
HELDIG 2017

Digital Language Typology (DLT)

Academy of Finland project

Computer-based platform to assess the
structurally manifested family relationships
within **any set of languages**
with appropriate
large digital textual and speech material

- Automatically discovery language family relationships
- Analysis of language data
- Without relying on language-specific tools, resources, etc.

Digital Language Typology (DLT)

Academy of Finland project

Computer-based platform to assess the
structurally manifested family relationships
within **any set of languages**
with appropriate
large digital textual and speech material

- Automatically discovery language family relationships
- Analysis of language data
- Without relying on language-specific tools, resources, etc.

Digital Language Typology (DLT)

Academy of Finland project

Computer-based platform to assess the
structurally manifested family relationships
within **any set of languages**
with appropriate
large digital textual and speech material

- Automatically discovery language family relationships
- Analysis of language data
- Without relying on language-specific tools, resources, etc.

Digital Language Typology (DLT)

Academy of Finland project

Computer-based platform to assess the
structurally manifested family relationships
within **any set of languages**
with appropriate
large digital textual and speech material

- Automatically discovery language family relationships
- Analysis of language data
- Without relying on language-specific tools, resources, etc.

Low-Resourced Languages

- Standard language processing tools not available
  part-of-speech taggers, parsers, semantic role labellers
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - But raw data might be available
- E.g. many Uralic languages
  - Not many speakers
    E.g. Nganasan, ~100 speakers

Low-Resourced Languages

- Standard language processing tools not available
  part-of-speech taggers, *parsers*, *semantic role labellers*
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - But raw data might be available
- E.g. many Uralic languages
  - Not many speakers
  - E.g. Ingrian: ~100 speakers

Low-Resourced Languages

- Standard language processing tools not available
  part-of-speech taggers, *parsers*, *semantic role labellers*
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - **But** raw data might be available
- E.g. many Uralic languages
  - Not many speakers
  - E.g. Nganasan: ~1000 speakers

Low-Resourced Languages

- Standard language processing tools not available
  part-of-speech taggers, *parsers*, *semantic role labellers*
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - **But** raw data might be available
- E.g. many Uralic languages
  - Not many speakers
  - E.g. Ikpeng has ~370 speakers

Low-Resourced Languages

- Standard language processing tools not available
  part-of-speech taggers, *parsers*, *semantic role labellers*
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - **But** raw data might be available
- E.g. many Uralic languages
  - Not many speakers

E.g. Nganasan, ∼100 speakers

Low-Resourced Languages

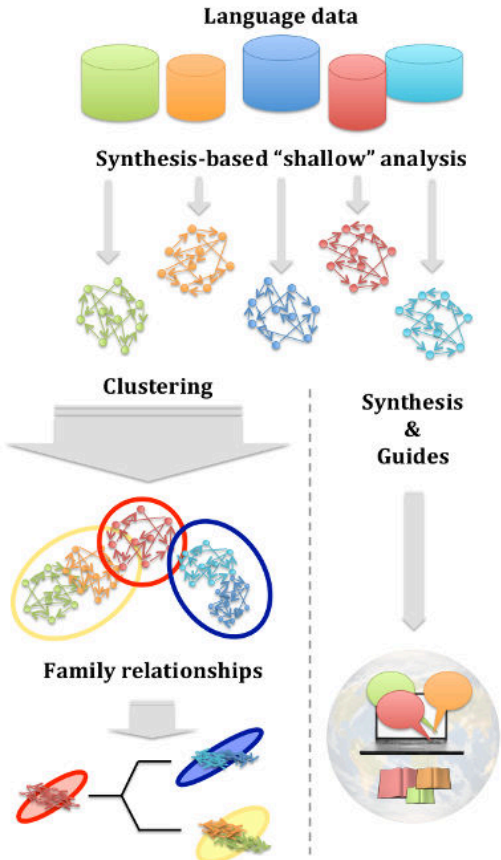
- Standard language processing tools not available
  part-of-speech taggers, parsers, semantic role labellers
- Not much annotated data
  - E.g. large corpora of expert linguistic analyses
  - **But** raw data might be available
- E.g. many Uralic languages
  - Not many speakers
    E.g. Nganasan, ∼100 speakers

Language Typology

- Grouping of languages according to their characteristics
- Mapping multi-dimensional space of:
  - similarities, differences
  - influence of contact
  - syntax, morphology
  - phonotactics, prosody

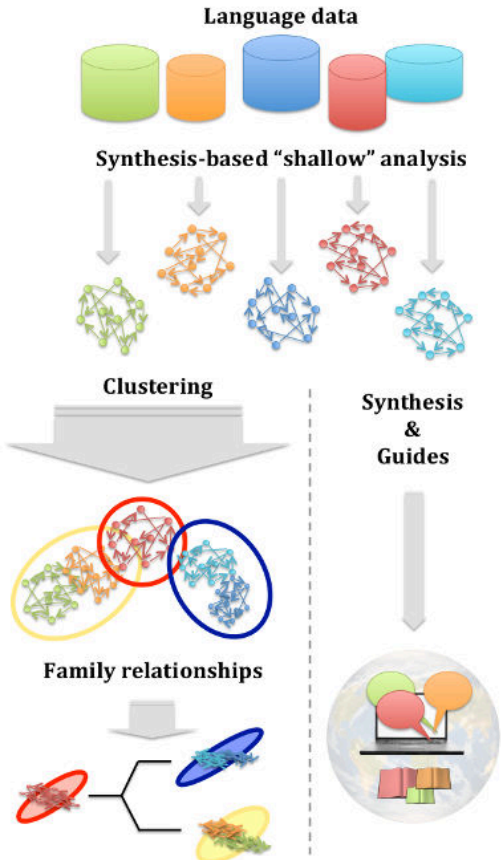
DLT:
"Mining from the surface to the core"

- Instead of labour-intensive study of languages, see what we can learn automatically from surface
- Text, speech recordings, transcribed speech
- Working with Martti Vainio's group on speech data
- Also looking at text input



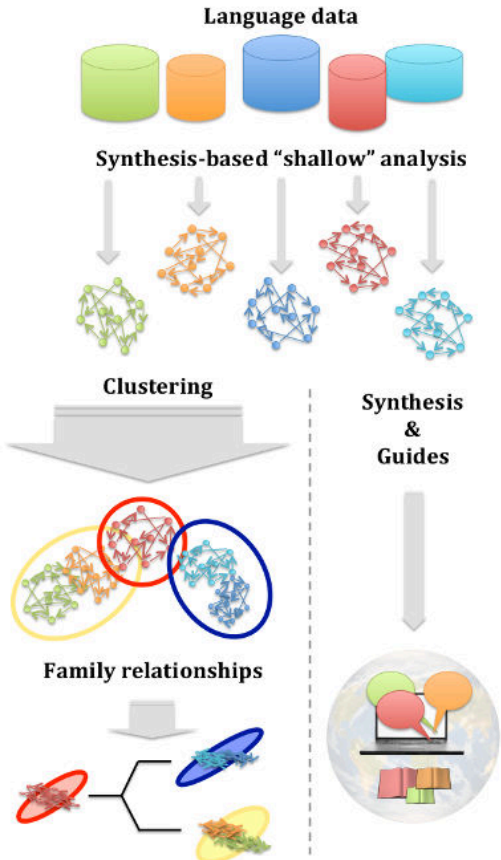
DLT:
"Mining from the surface to the core"

- Instead of labour-intensive study of languages, see what we can learn automatically from **surface**
- Text, speech recordings, transcribed speech
- Working with Martti Vainio's group on speech data
- Also looking at text input



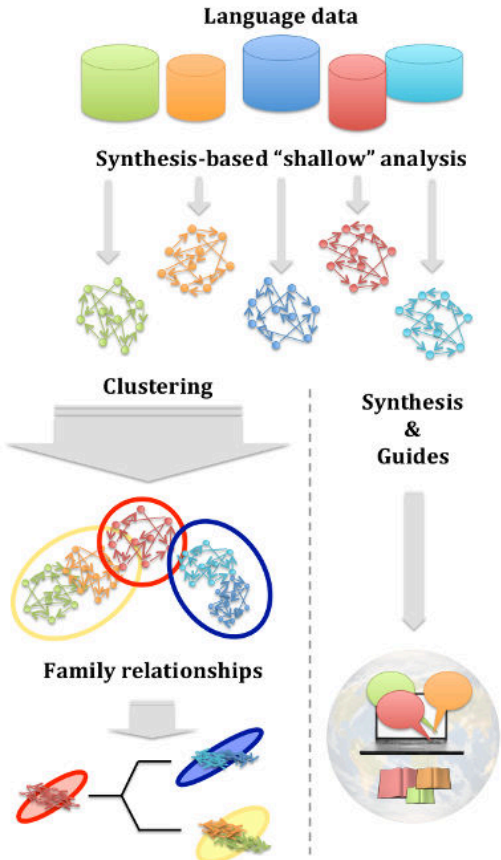
DLT:
"Mining from the surface to the core"

- Instead of labour-intensive study of languages, see what we can learn automatically from **surface**
- Text, speech recordings, transcribed speech
- Working with Martti Vainio's group on speech data
- Also looking at text input



DLT:
"Mining from the surface to the core"

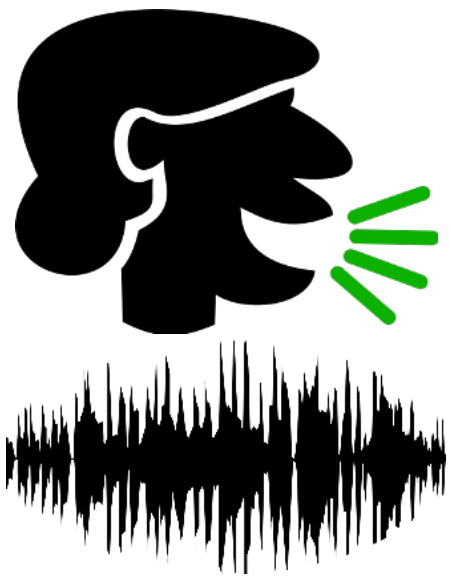
- Instead of labour-intensive study of languages, see what we can learn automatically from **surface**
- Text, speech recordings, transcribed speech
- Working with Martti Vainio's group on speech data
- Also looking at text input



Current work in Discovery Group

Question:

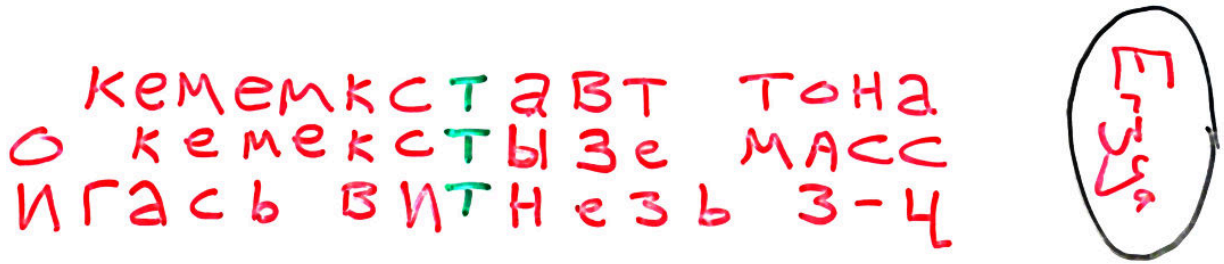
How much can we discover automatically about languages we know nothing about to start with?



nɛsdə pɒŋt pɒ dʔɑwsɒdənən

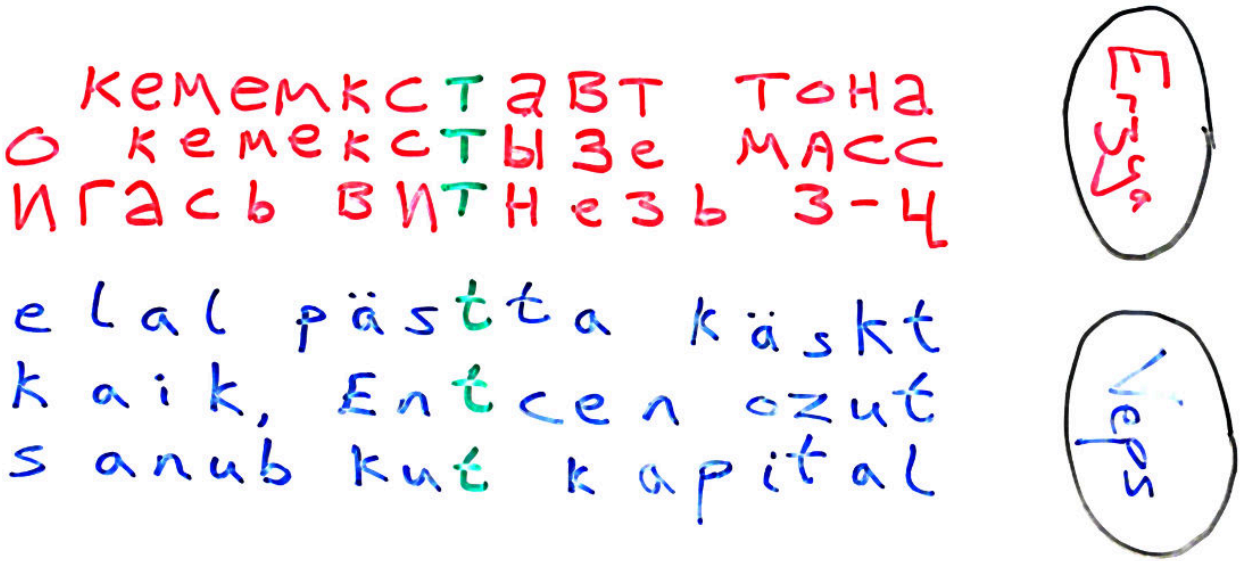
nɛsta pəŋkt poː føːrədrɑːŋnɪŋslɪstan

Current work in Discovery Group



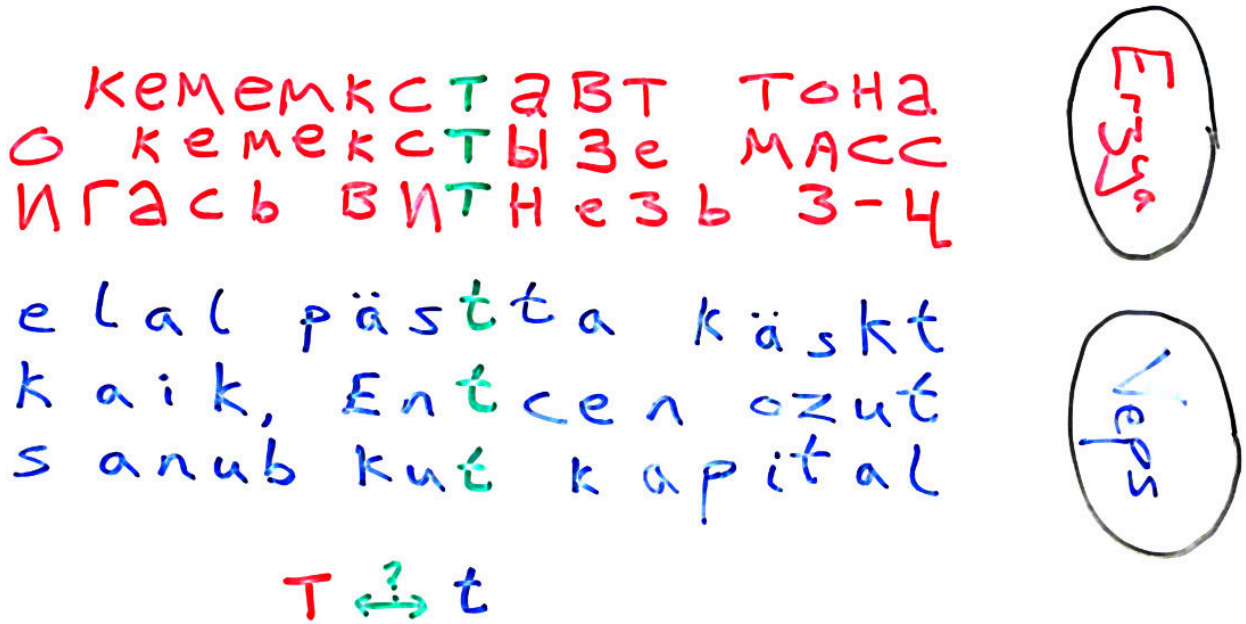
Find similarities between **characters/sounds** of different languages, in terms of how they are used

Current work in Discovery Group



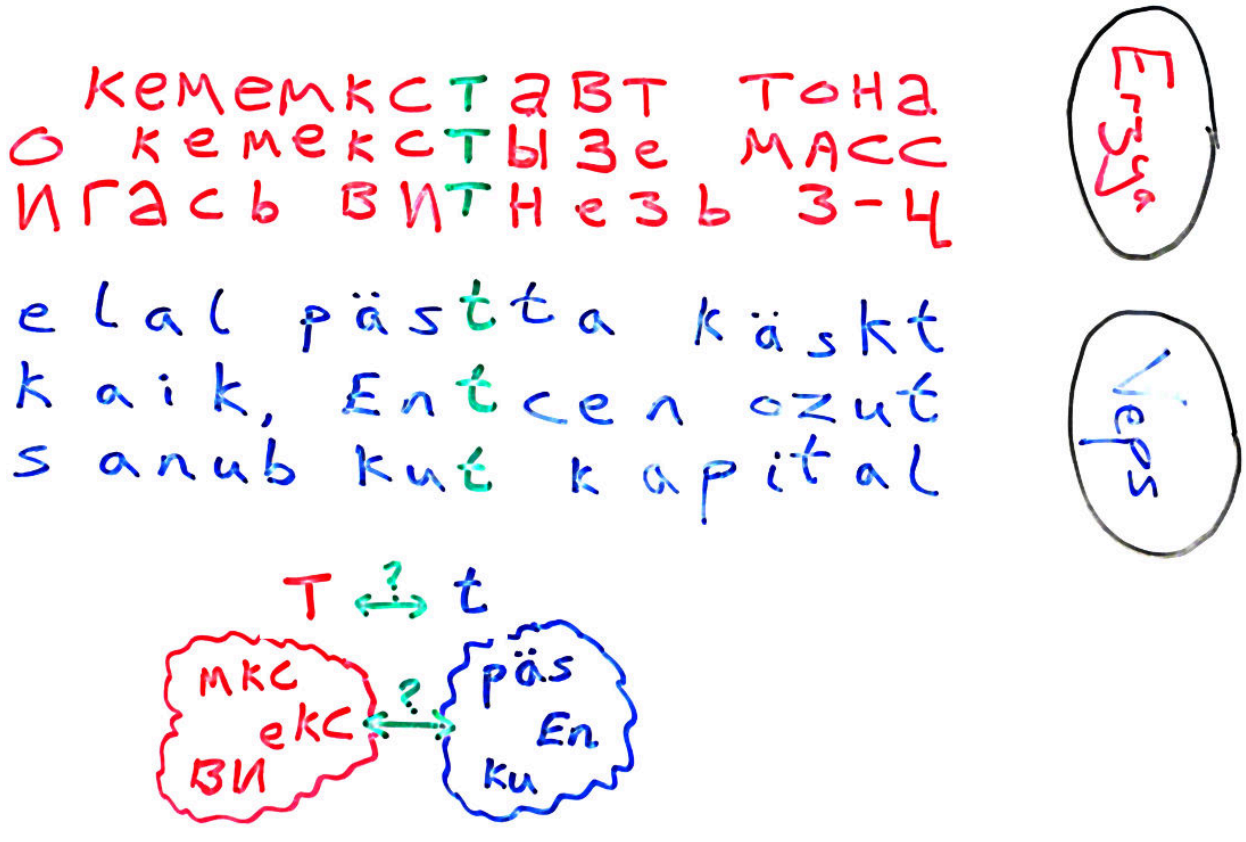
Look at **contexts** they appear in

Current work in Discovery Group



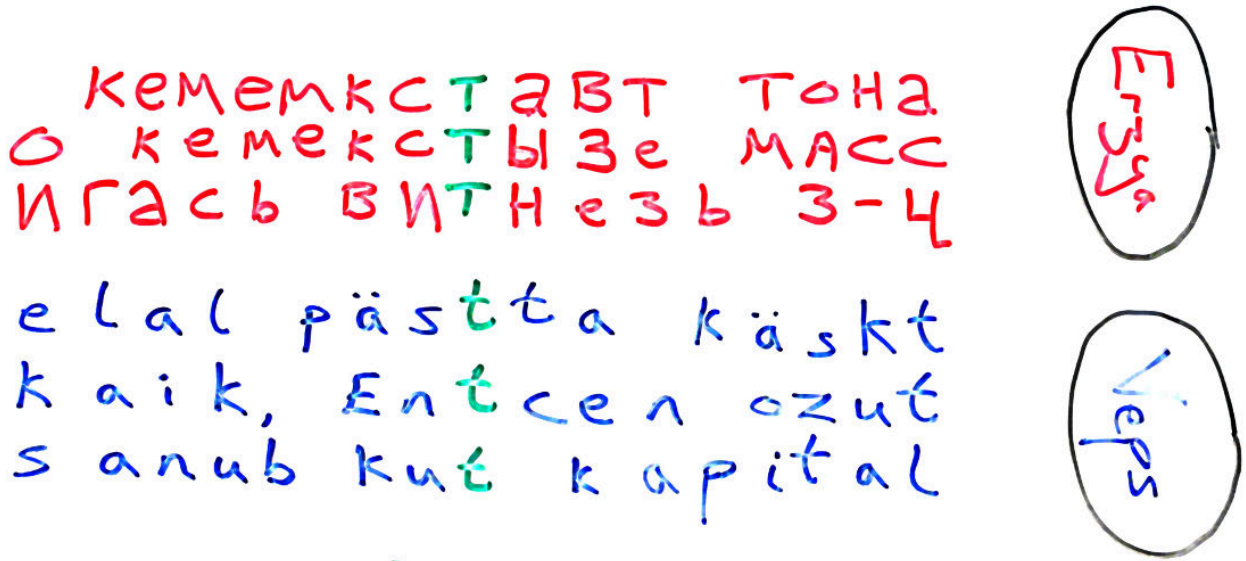
Look at **contexts** they appear in

Current work in Discovery Group



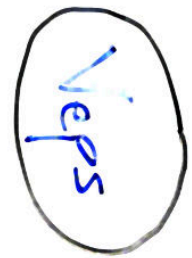
Difficulty: contexts specific to each language, can't compare

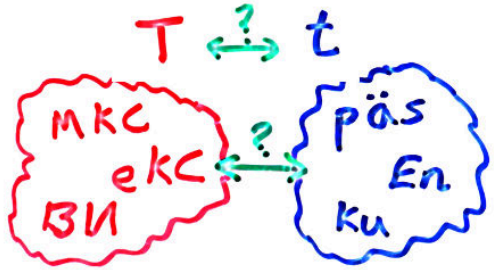
Current work in Discovery Group



Novel **neural network** method finds
common patterns of co-occurence
Promising results

Current work in Discovery Group





тӣдтэ ⟷ tietää

Can possibly use to **identify related words** where languages

- use different sounds to express them
- or write them differently (e.g. different scripts)