HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Computational methods for forming a nation-wide toponymic overview

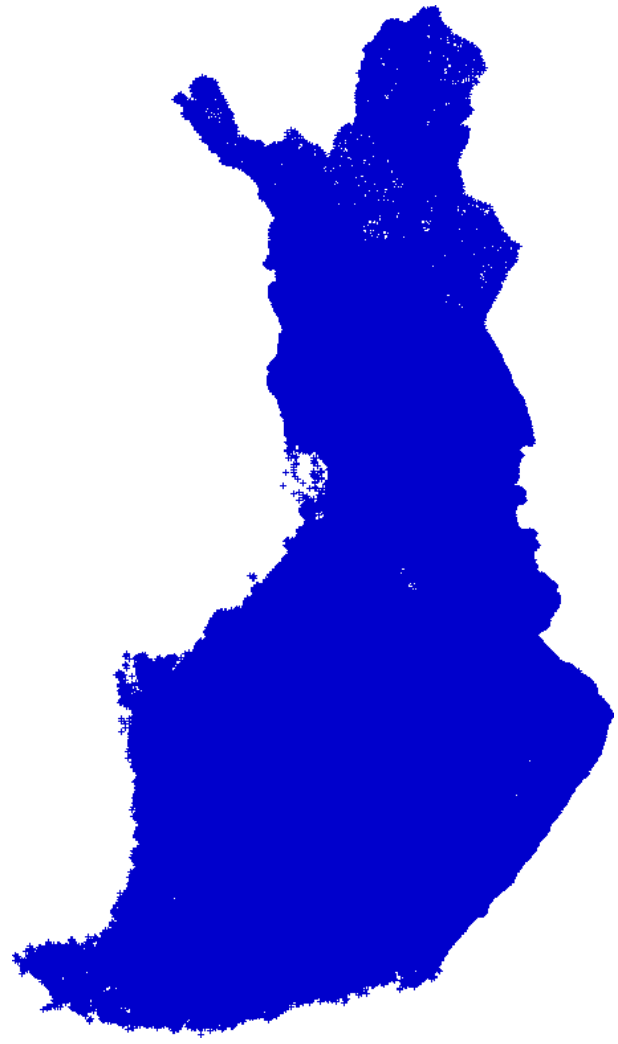
Antti Leino ‹antti.leino@cs.helsinki.fi›

28th November 2006

Introduction

So many names, so little time

- Lots of place names in a country
  - Finnish 1:20 000 Basic Map has
    - c. 800 000 named places
    - c. 360 000 different names
- Not feasible to study 360 000 distribution maps
- How to present the overall variation?



Introduction

What can we do?

- Data mining
  - Sub-field of computer science
  - Goal: find interesting new information in large collections of data
- Here: some examples of what can be done
  - Visualisation
  - Computational analysis
- Choice of tools depends on the data

Introduction

Languages in Finland

- Two official languages
  - Finnish (91.64 %)
  - Swedish (5.50 %)
- Five semi-official languages
  - Sámi languages (0.03 %)
    - Northern Sámi
    - Enare Sámi
    - Skolt Sámi
  - Romany
  - Finnish sign language

- Finnish, Swedish and the Sámi languages are used on maps

Introduction

Getting to know the data

- Place Name Register
  - Kept by the National Land Survey
  - Part of the map-making process

Language	Names	Places
Finnish	303 626	717 747
Swedish	48 319	74 726
Northern Sámi	4 115	4 529
Enare Sámi	3 306	3 774
Skolt Sámi	141	148
Total	359 507	800 924

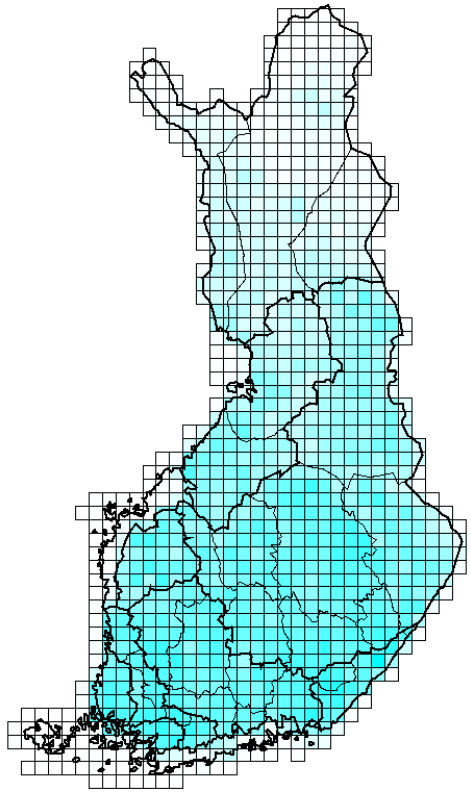
Languages in Toponyms

Visualisation

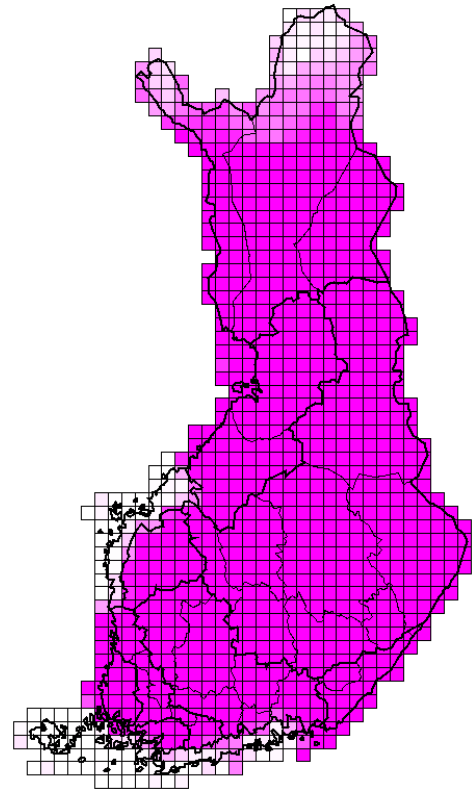
- Simple way to visualise the different languages:
  - Divide the contry into 20×20 km squares
  - Count the place names in each language in each square
  - Display these on a map
- Variation: how many % of the square's toponyms are in each of the languages?
- Computationally easy, good first step

Languages in Toponyms

Finnish



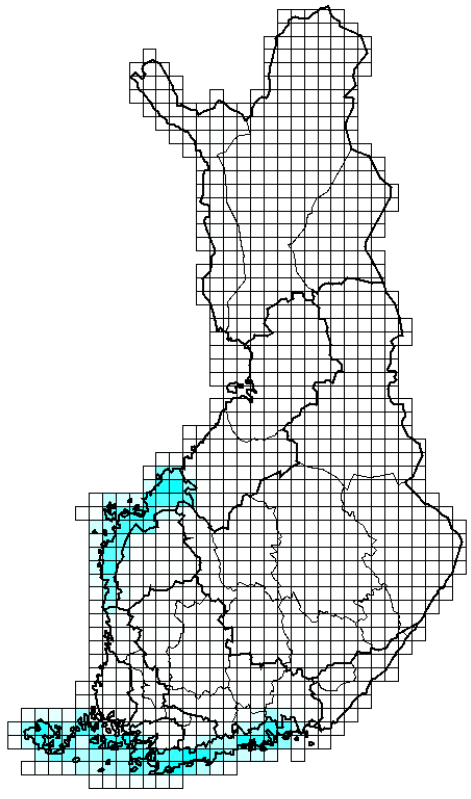
Absolute
max=2246



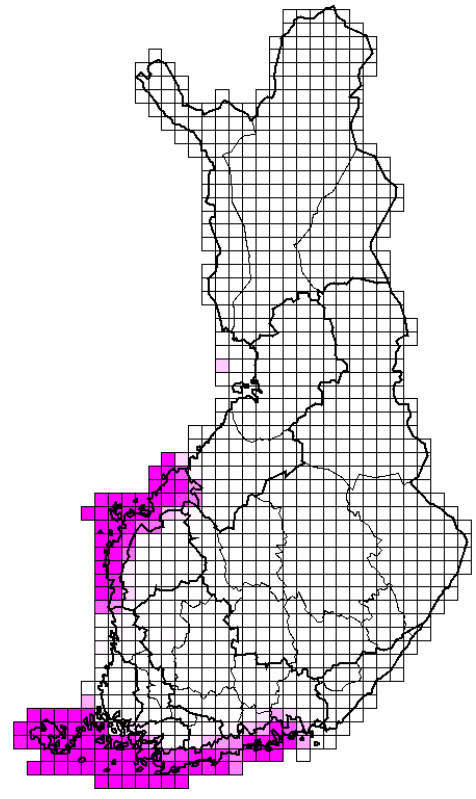
Relative
max=100 %

Languages in Toponyms

Swedish



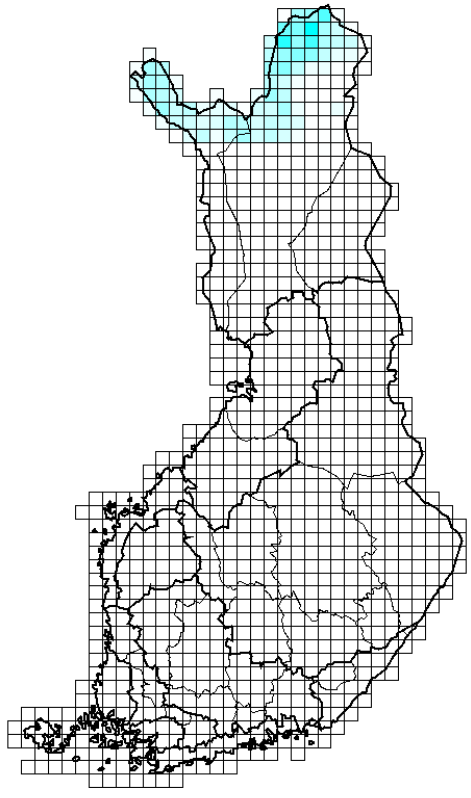
Absolute
max=1597



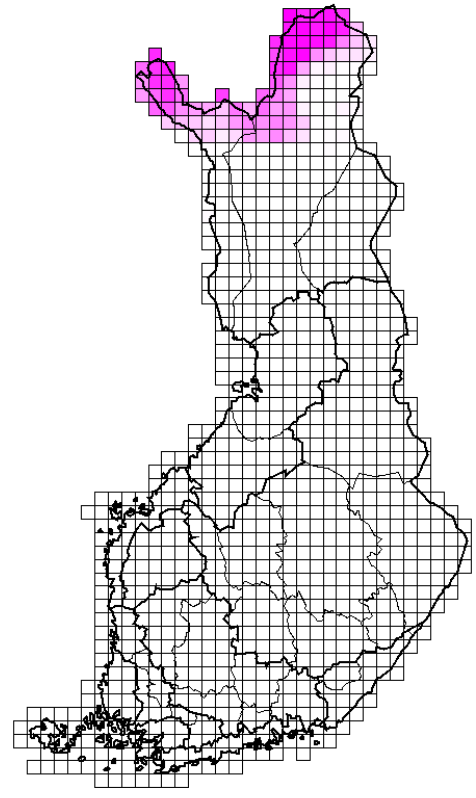
Relative
max=100 %

Languages in Toponyms

Northern Sámi



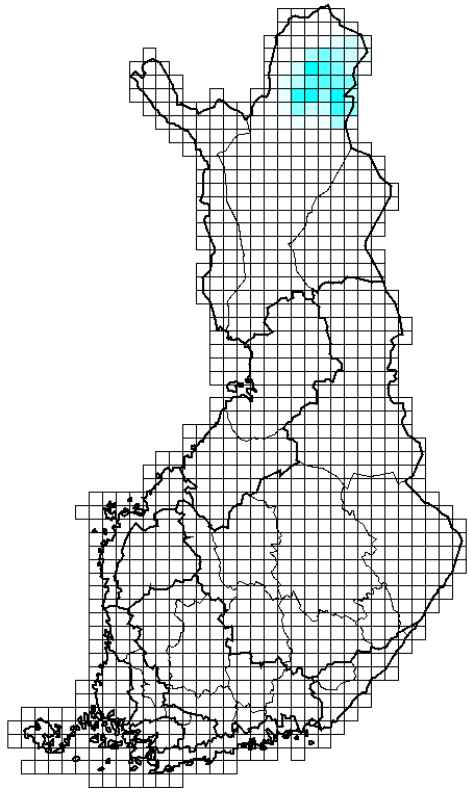
Absolute
max=234



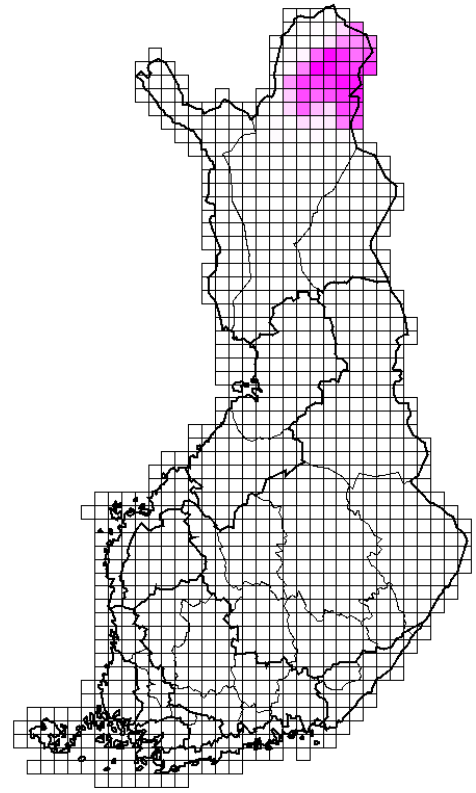
Relative
max=100 %

Languages in Toponyms

Enare Sámi



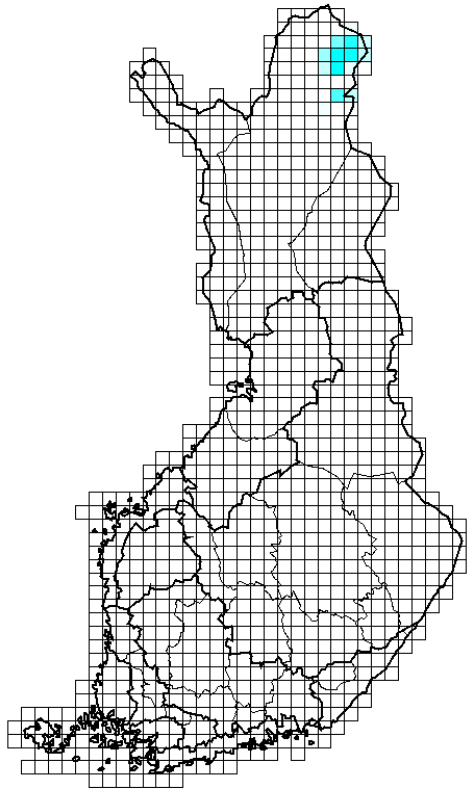
Absolute
max=285



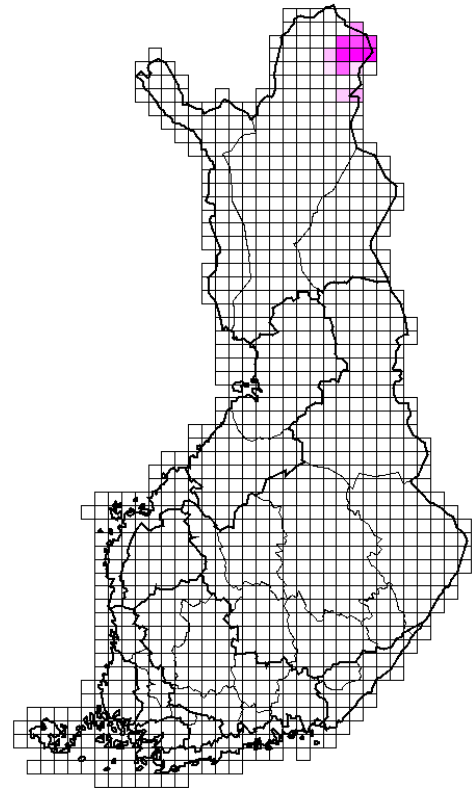
Relative
max=65 %

Languages in Toponyms

Skolt Sámi



Absolute
max=23



Relative
max=14 %

Languages in Toponyms

So what?

- Finnish is a clear majority language
- This is reflected in place names

- So few Sámi toponyms that a more thorough onomastic overview is not meaningful
- With Swedish such an overview could be useful

- Finnish names used here to illustrate further methods

Variation in Names

- Goal: summarise most notable aspects of variation
- Most common names in different regions
  - Computationally and conceptually easy
  - Not always very informative
- Underlying components that explain the variation
  - Sophisticated statistical / computational methods
  - Not always intuitive
  - Can be more informative

Variation in Names

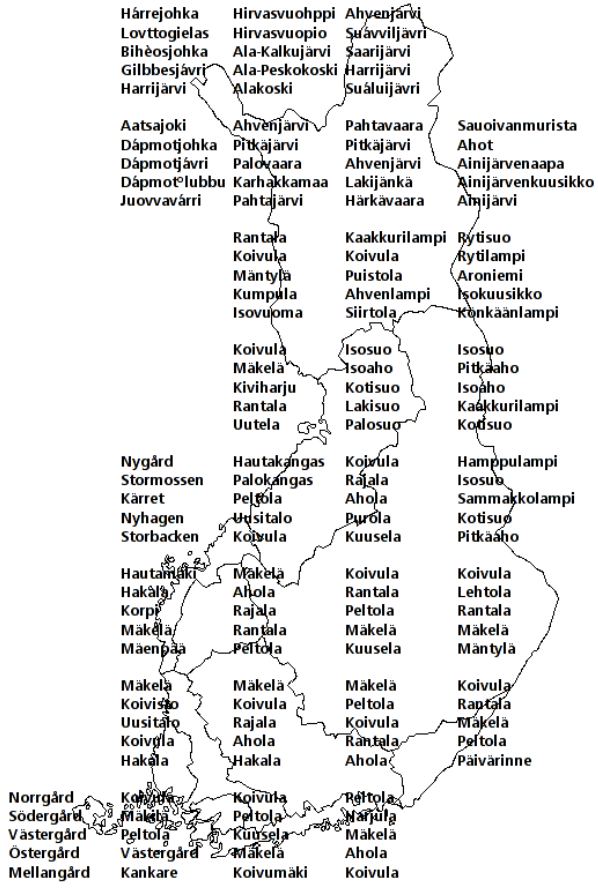
Most Common Names

- Divide country to e.g. 150×150 km squares
- Write on map the most common names

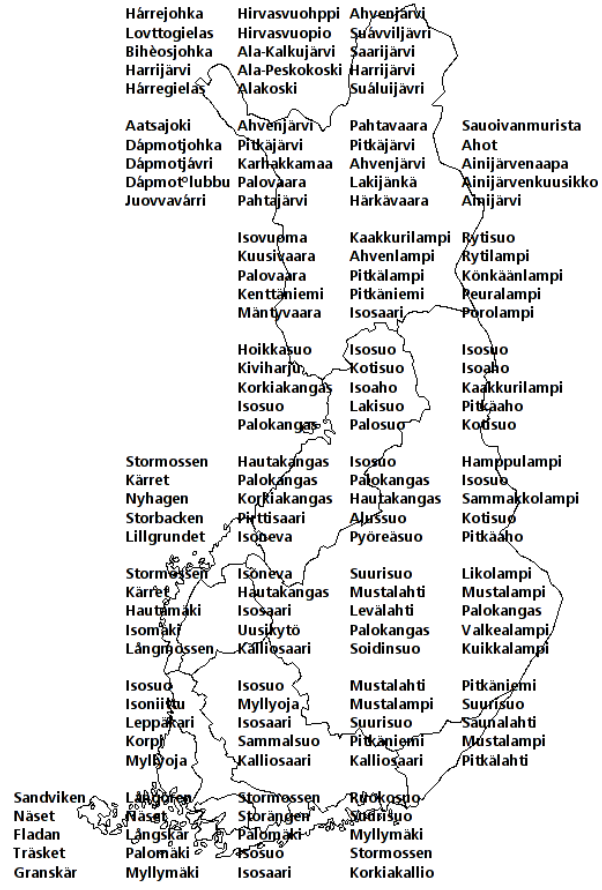
- Variant: name elements instead of complete names
  - Finnish names often consist of two parts, e.g. *Mustalampi*: *musta* 'black' + *lampi* 'pond'
  - Last elements shows the type of place
  - First part describes / identifies the place

Variation in Names

Most Common Names



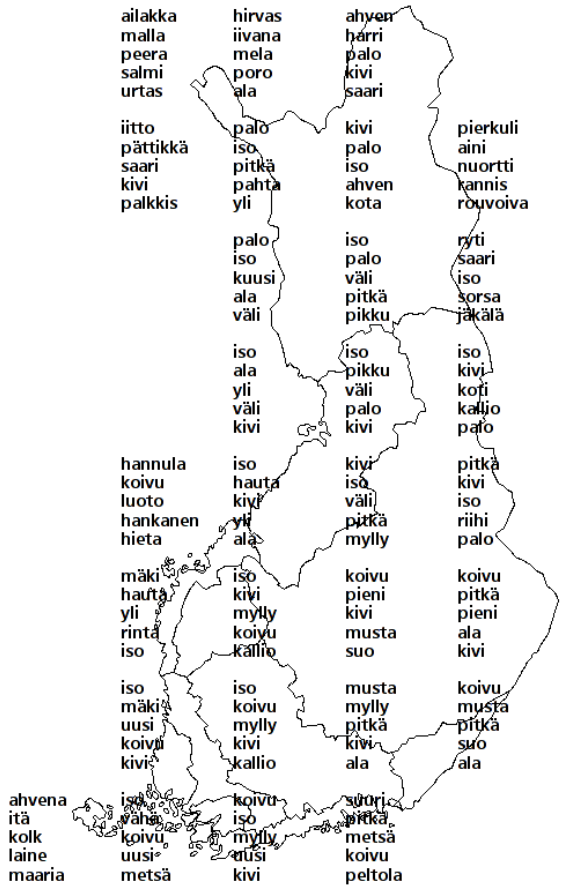
All



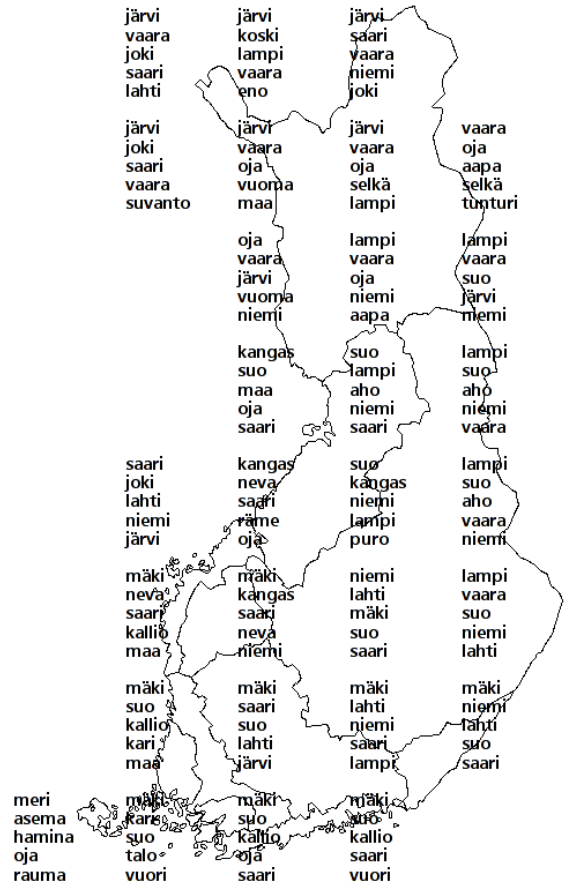
Natural Features

Variation in Names

Most Common Name Elements



First



Last

Onomastic Regions

How to find?

- Goal: present regional toponymic variation concisely
- Concise: at most 10–20 maps
- Two main alternatives
  - Clustering
  - Component / Factor Analysis

Onomastic Regions

Clustering

- Overall goal: divide the data into groups (≈ regions) so that
  - Data items (≈ municipalities / grid cells) in the same cluster as similar as possible
  - Those in different clusters as different as possible
- Problematic for linguistic variation in general
  - Variation is gradual, no clear borders between regions
- Especially so for toponyms

Onomastic Regions

Component and Factor Analysis

- Goal: find factors that explain the overall variation
- Analogy: traditional dialectology
  - Determine dialect borders by combining individual isoglosses
  - The isoglosses are weighted: some features are considered more important than others
- Here, the same thing but automatically
  - Distributions of different toponyms are combined
  - The weight of each toponym is determined so that the overall division is maximally clear

Onomastic Regions

Non-negative Matrix Factorisation

- Designed for non-negative data
  - This applies here: the number of names in a region ≥ 0
- Pretty much the same results as with traditional Factor Analysis
- Computationally much faster
- By no means the only method available
  - Use one you (or your pet data analyst) are comfortable with

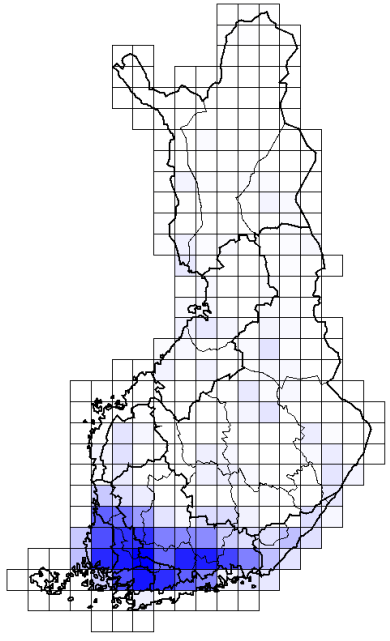
Onomastic Regions

Regions in Finland

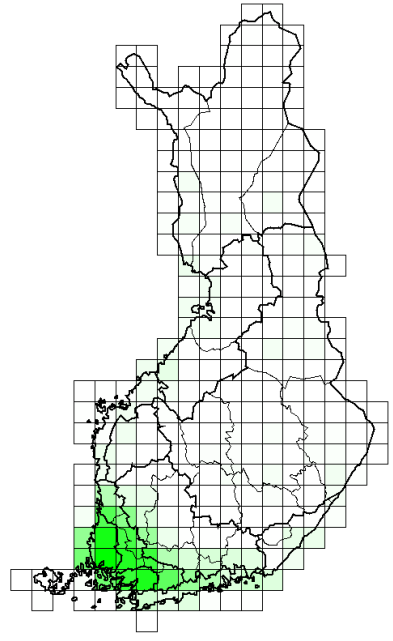
- NMF applied to three different data sets
  - All names on the 1:20 000 Basic Map
    name ≡ (written form, type of place, language)
  - First parts of at most two-part names in Finnish: **Musta**lampi
  - Last parts of at least two-part names in Finnish: Musta**lampi**
- 40 × 40 km squares, occurrence of names in a square as 1/0
- Factors shown as maps
- Result: ‘regions’ as diffusion patterns

Onomastic Regions

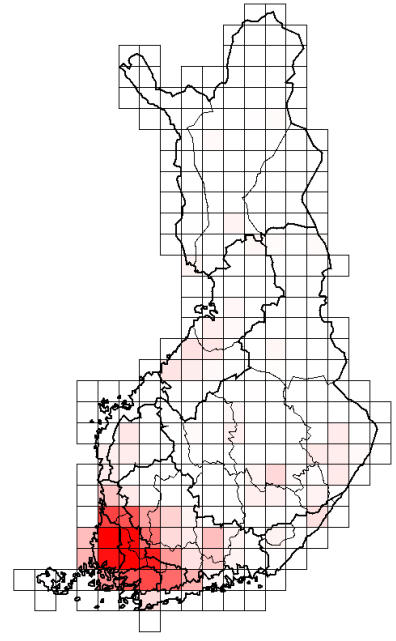
Finland Proper



All
names



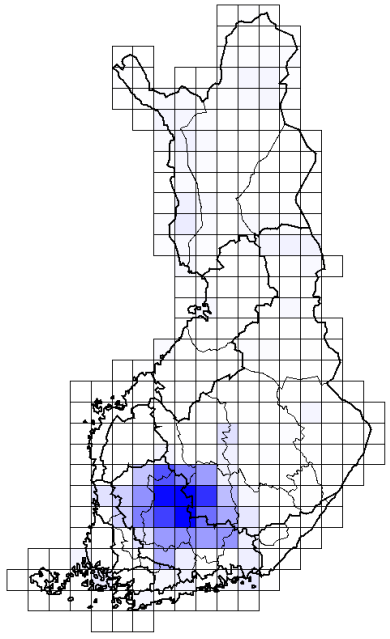
Finnish
first parts



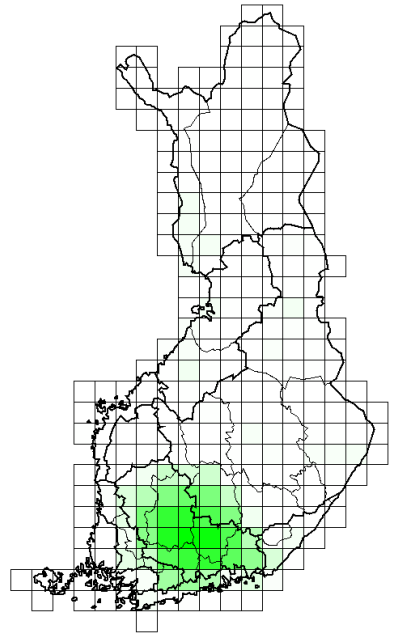
Finnish
last parts

Onomastic Regions

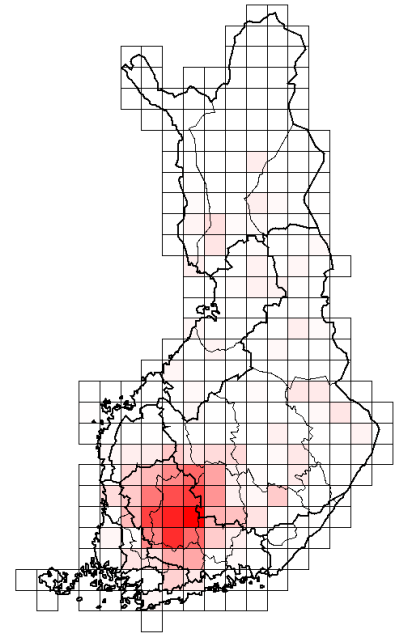
Tavastia



All names



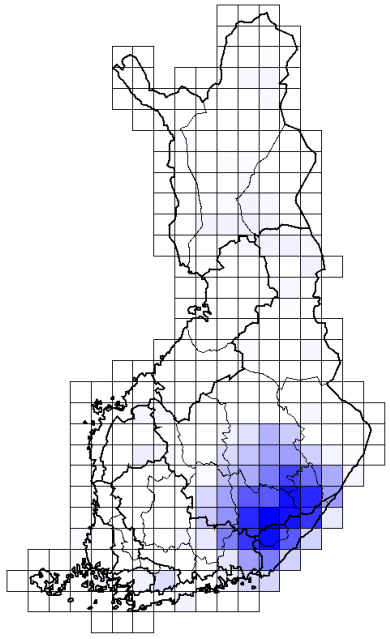
Finnish first parts



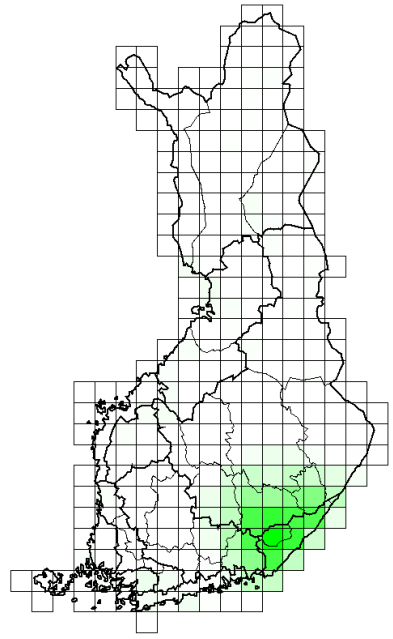
Finnish last parts

Onomastic Regions

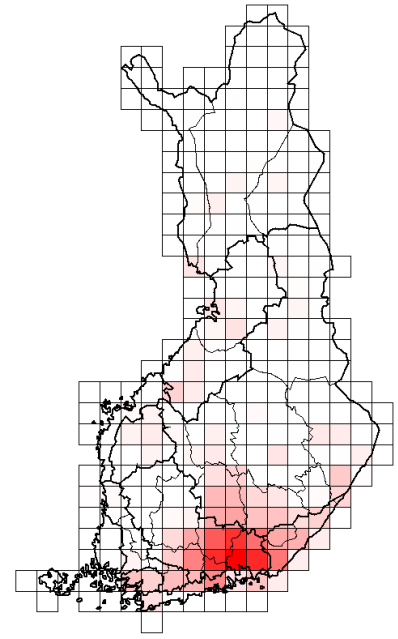
Southern Carelia



All names



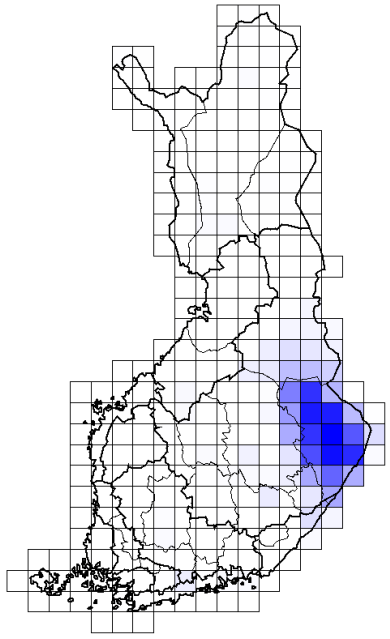
Finnish first parts



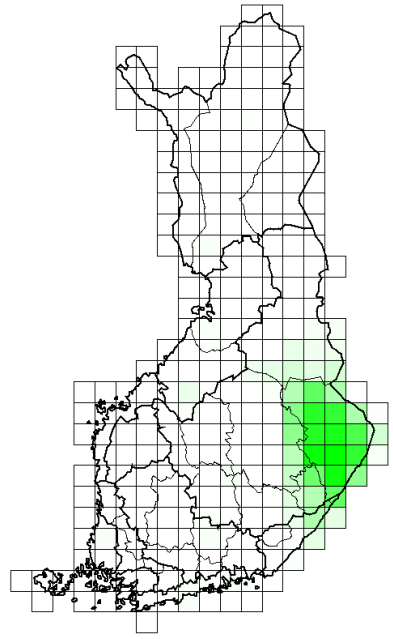
Finnish last parts

Onomastic Regions

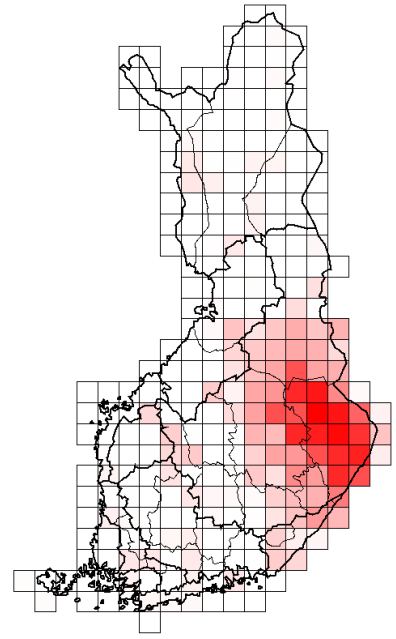
Northern Carelia



All names



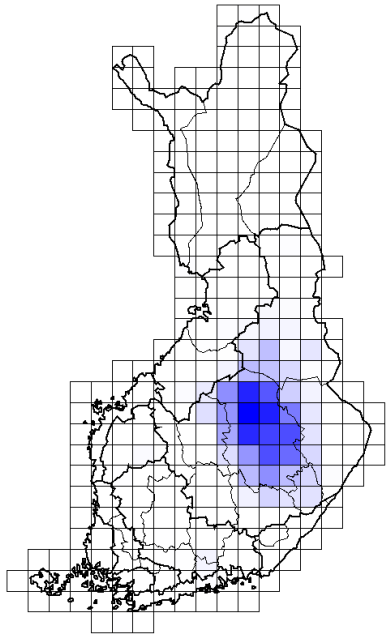
Finnish first parts



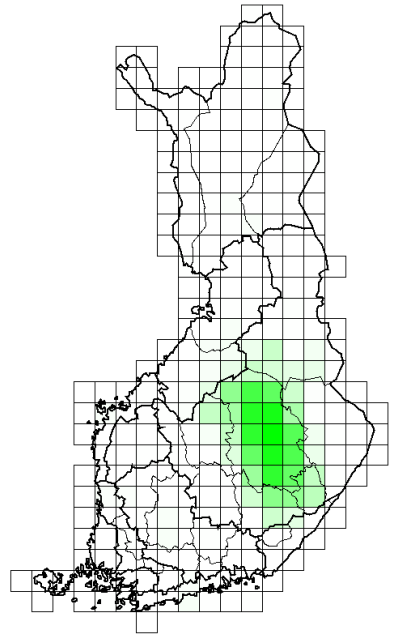
Finnish last parts

Onomastic Regions

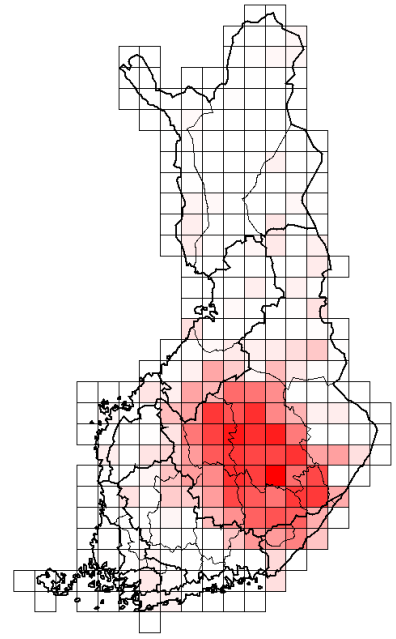
Savonia



All names



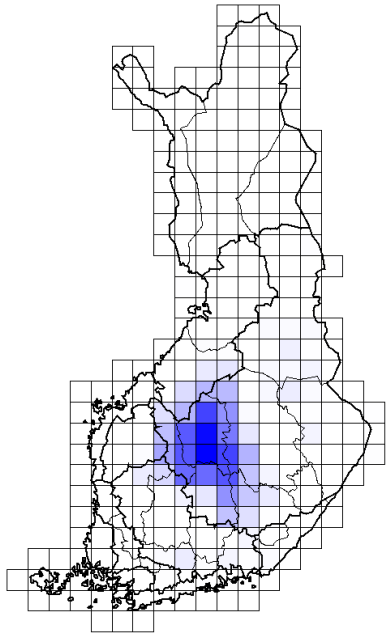
Finnish first parts



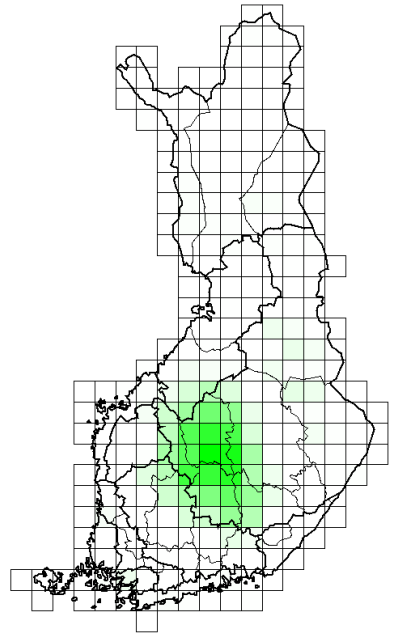
Finnish last parts

Onomastic Regions

Western Savonia / old Tavastian wilderness



All names

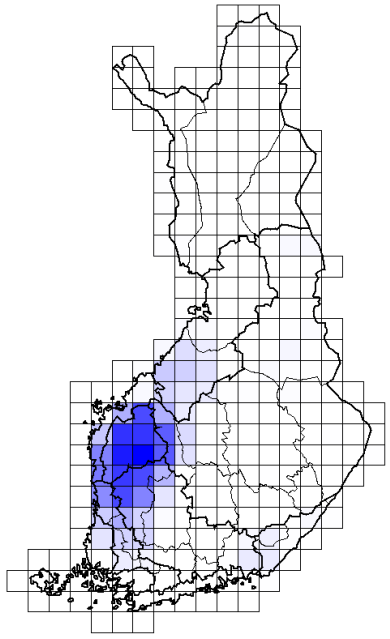


Finnish first parts

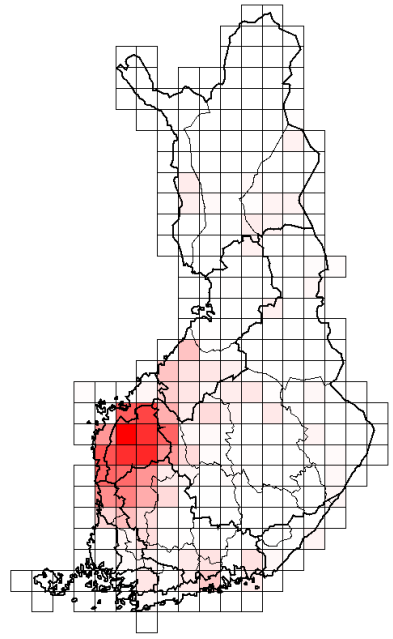
Finnish last parts

Onomastic Regions

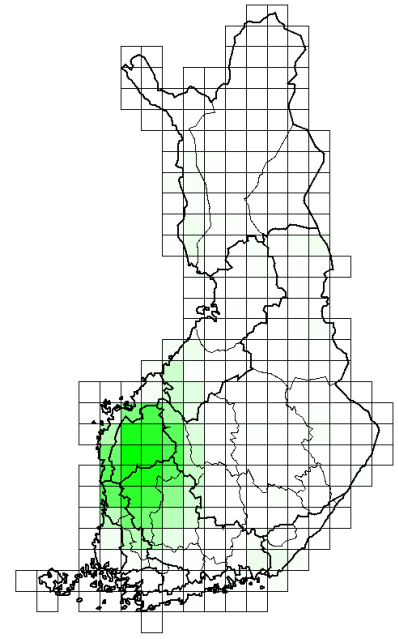
Southern Ostrobothnia



All names



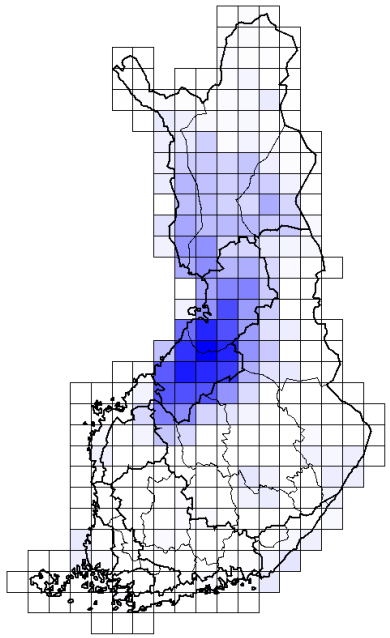
Finnish first parts



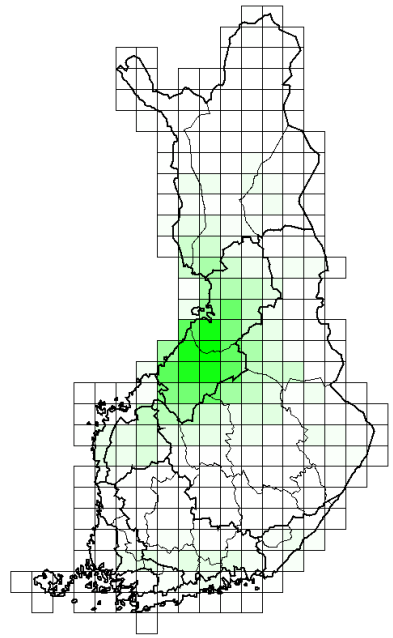
Finnish last parts

Onomastic Regions

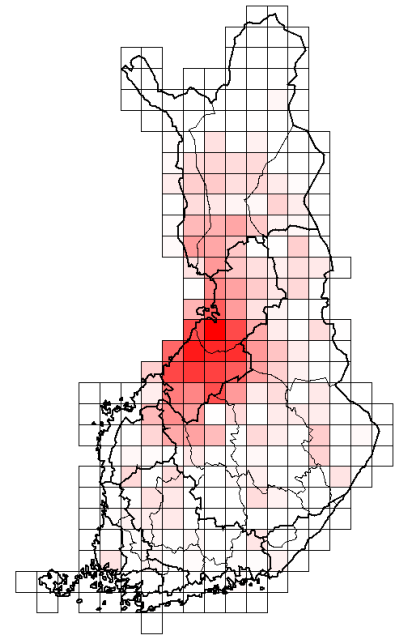
Central / Northern Ostrobothnia



All
names



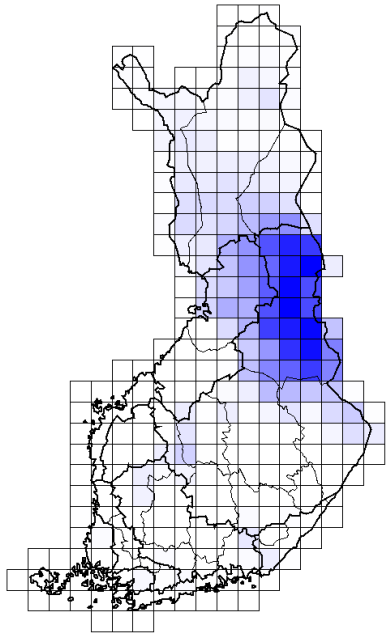
Finnish
first parts



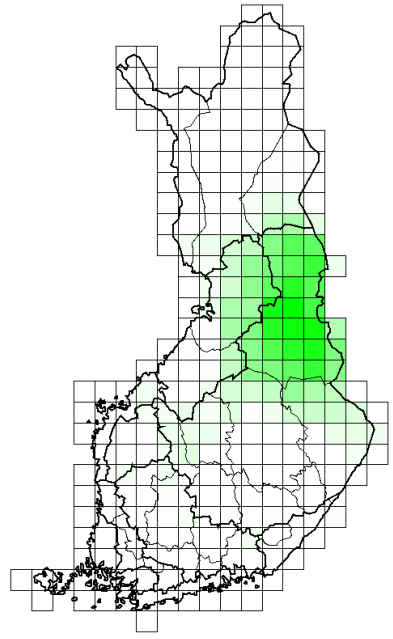
Finnish
last parts

Onomastic Regions

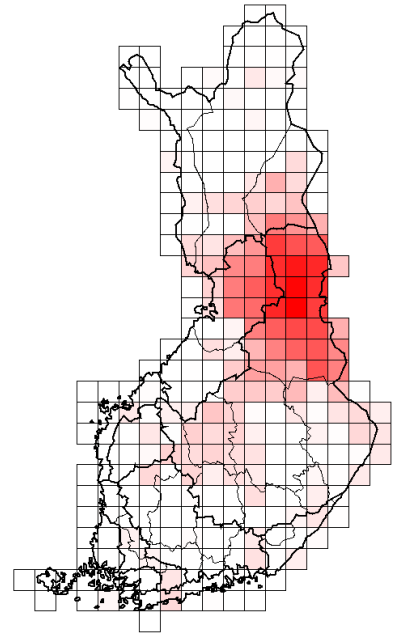
Kainuu



All names



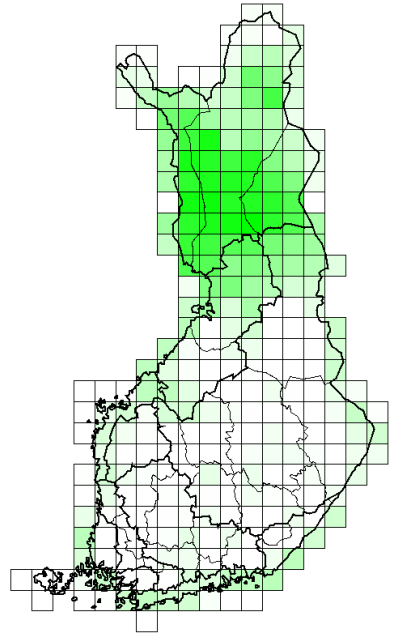
Finnish first parts

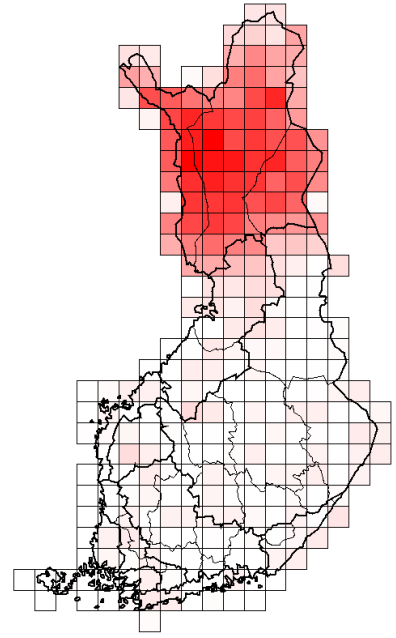


Finnish last parts

Onomastic Regions

Lapland





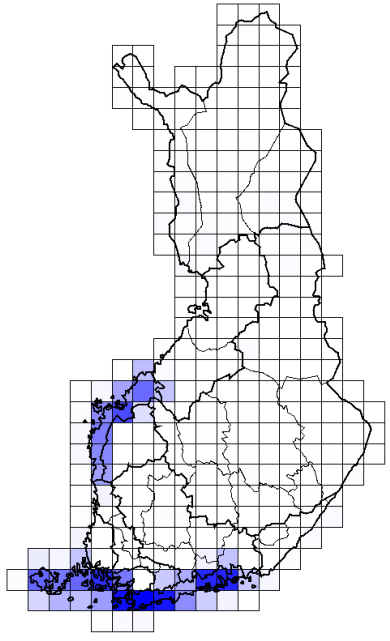
All names

Finnish first parts

Finnish last parts

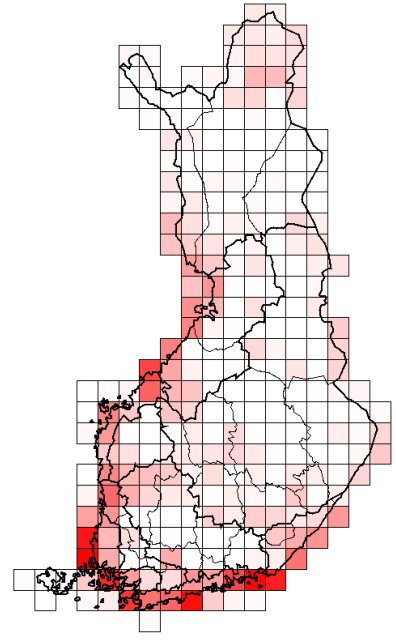
Onomastic Regions

Swedish-language coast



All
names

Finnish
first parts



Finnish
last parts

Summary

- Some processing is required to get a one-glance overview of a large onomastic corpus
- There are various computational methods that can be used
  - Name counts for grid cells
  - Most common names / elements in grid cells
  - Factor analysis
  - Plenty of others
- Visualisation in the form of maps
- Choice of tools depends on the goals of the onomastic study

Thank you