

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



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?





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



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



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	48319	74726	
Northern Sámi	4115	4 529	
Enare Sámi	3 306	3774	
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







Absolute max=2246

Relative max=100 %







Absolute max=1597

Relative max=100 %



Languages in Toponyms Northern Sámi





Absolute max=234

Relative max=100 %







Absolute max=285

Relative max=65 %







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

	Hárrejohka	Hirvasvuohppi	Ahyenjarvi	
	Lovttogielas	Hirvasvuopio	Salavviljavn	
	Bihèosjohka	Ala-Kalkujarvi	Saarijarvi	5
	Gilbbesjävri	Ala-Peskokoski	Harrijarvi /	
	Harrijarvi 🖓 🗌	Alakoski	Suáluijavri	
	. \	1 ~ ~		
	Aatsajoki 📏	Ahvenjarvi	Pahtavaara /	Sauoivanmurista
	Dápmotjohka	Pitkajarvi	Pitkajarvi 🔪	Ahot
	Dápmotjávri	Paloxaara	Ahvenjarvi 🛰	Ainijärvenaapa
	Dápmot ^o lubbu	Karhakkamaa	Lakijanka	Ainijärvenkuusikko
	Juovvavarri	Pahtajarvi	Harkavaara	Ahjijärvi
		4		1
		Rantala	Kaakkurilampi	Rýtisuo
		Koivula	Koivula	/Rytilampi
		Mantyfa	Puistola	Aroniemi
		Kumpala	Ahvenlampi	heokuusikko
		Isovujoma	Siirtola ~~~~~	Kènkaanlampi
			~~	. \
		Koivula	Isosuo 4	Isosiuo
		Makela"	-Isoaho	Pitkaaho
		Kiviharju "	Kotisuo (Isoaho
		Rantala j	Lakisuo 👌	Kaékkurilampi
		Uutela 👧	Palosuo	Kodsuo
	Numbed	Hutter frage	Kohola	Hammond amou
	Stormorron	Balok maar	P.data	Icorup
	Variat	Palthia	Abola	Sammakkolamni
	Mahamm	Haritalo	Durfile	Koticue
	Stochaster 2	Kabada	Kuncela	nith him
	Sumacken	Rugadia /	Kuuseia	riteaalo
	Haut Link 2	TARA AND A	Kohula	Kobula
	Hakata	Ahola	Rantala	Lehtola
	Korol	Rafala	Peltola	Rantala /
	MAN AND	Rantala	Makala	Makela
	Mannaka	Ridfol a	Kuurala	Mantela
	17/	a concella	Kuussia	
	Makela	Makela	Makela	Koivula
	Koivisko 🔿	Koivula 3	Peltola	Raptala
	Uusitato	Rajala	Koivula /-	Makela
	Koivala	Ahola 😽	Rantela	Peltola
	Hakala	Hakala	Ahola	Paivarinne
	1 2 1	٤	5/	
Norrgård	Keikhthen	Kolvala	'Réltola	
Södergård 2	Maleta Z	Peltoja	Marjalla	
Vastergård 32	Peltola 2	Kuusela-1 1000	Makela	
Ostergård	Vastergard 2	Makela	Ahola	
Mellangård	Kankare	Koivumaki	Koivula	

	Hárrejohka	Hirvasvuohppi	Ahyenjarvi	
	Lovttogielas	Hirvasvuopio	Suavviljavn	
	Bihèosjohka	Ala-Kalkujarvi	Saarijarvi	5
	Harrijarvi	Ala-Peskokoski	Harrijarvi /	
	Harrenielas	Makoski	Suáluijavri	
	management	1	2	
	Aatsajoki 🛝	Ahvenjarvi	Pahtavaara /	Sauoivanmurista
	Dápmotjohka	Pitkäjärvi	Pitkajarvi 🔪	Ahot
	Dápmotjávri	Karhakkamaa	Ahvenjarvi	Ainijärvenaapa
	Dapmot ^o lubbu	Palovaara	Lakijanka	Ainijärvenkuusikko
	Juowawarri	Pahtajarvi	Harkavaara	Amijarvi
		ζ		1°
		Isovuema	Kaakkurilampi	Rýtisuo
		Kuusiyaara	Ahvenlampi	Rytilampi
		Palovaara	Pitkalampi	Konkaanlampi
		Kenttaniemi	Pitkaniemi	Reuralampi
		Mantiyyaara	Isosaari	Porolampi
			~~~	( .
		Hoikkasuo	Tisosuo L	Isosiuo
		Kivihariin	Kotisuo	Isoabo
		Korkiakangas	Isoaho	Kaakkurilampi
		Isosuo	Lakisuo 🔾	Pitkaaho
		Palokanoan	Palosur	Kotisuo
	Stormossen	Hautakannas	Isomo	Hamnoulampi
	Karret	Palokannas	Palokannas	Isosuin
	Nyhagen	Korfiakannas	Hautakannas	Sammakkolamni
	Storbacken	Distlicanti	Alufano	Kotisue
	Lillorundet 2	Ischera (	Bröreksung	Ritkasho
	a set	5	1) or cabao	indiana
	Stormwsen C	Isoneva (	Suurisuo	Likolampi
	Karper	Hautakangas	Mustalahti	Mustalampi
	Hautimaki	Isosaari	Levalahti	Palokangas /
	Isomak	Unsilveto	Palokangas	Valkealampi
	Lånorndssen.	Kalliosaari	Soldinsuo	Kuikkalampi
	11/	2		7
	Isosud	Isosuo 🤇	Mustalahti	Pitkaniemi
	Isoniittu	Myllyoia 3	Mustalampi	Suurisuo
	Leppakari	Isosaari	Suurisuo /~	-Saunalahti
	Korpf	Sammalsuo 😽	Pitkaniemi	Mostalampi
	Mylipola	Kalliosaari	Kalliosaari	Pitkalahti
	1.1		5/	
Sandviken 🔿	Laborron/	Stormossen	Reckosto	
Naset 63	Baset 2	Storanger	Speringuo	
Fladan 3.8	Langskar 2	Palomaki-	Myllymaki	
Trasket	Palomaki 2	Tosuo	Stormossen	
Granskår	Myllymäki	Isosaari	Korkiakallio	
THE R PROPERTY AND A REPORT OF			A REAL PROPERTY AND A REAL	

#### 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: Mustalampi
  - Last parts of at least two-part names in Finnish: Mustalampi
- 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**









#### Onomastic Regions Southern Carelia





#### Onomastic Regions Northern Carelia









#### Onomastic Regions Western Savonia / old Tavastian

#### wilderness



Finnish last parts



### Onomastic Regions Southern Ostrobothnia





## **Onomastic Regions** Central / Northern Ostrobothnia



















#### 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





