Using Gabmap

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

LanCLiD2, April 30, 2015

Gabmap – what can it do?

Gabmap is a web-based application for dialectometry, the quantitative study of dialectal differences.

- Inspect, visualize comparative data in various levels of detail
- Clustering: dendrograms and the maps
- MDS
- Cluster validation using MDS
- Probabilistic clustering
- Relations between linguistic differences and geographic distances
- Finding items characteristic to a particular dialect group

The Data

Gabmap can analyze

- string data (typically phonetic transcriptions)
- numeric data (e.g., counts of lexical items)
- categorical data
- externally calculated difference data

Gabmap accepts Unicode text files in tabular format.

The Data

Gabmap can analyze

- string data (typically phonetic transcriptions)
- numeric data (e.g., counts of lexical items)
- categorical data
- externally calculated difference data

Gabmap accepts Unicode text files in tabular format.

	boter	**zout**
Aalsmeer	botər	zaut
Baardegem	botər	zat
Blankenberge	bøtər	zut
Coevorden	bœtər	sɔl̥t
...		

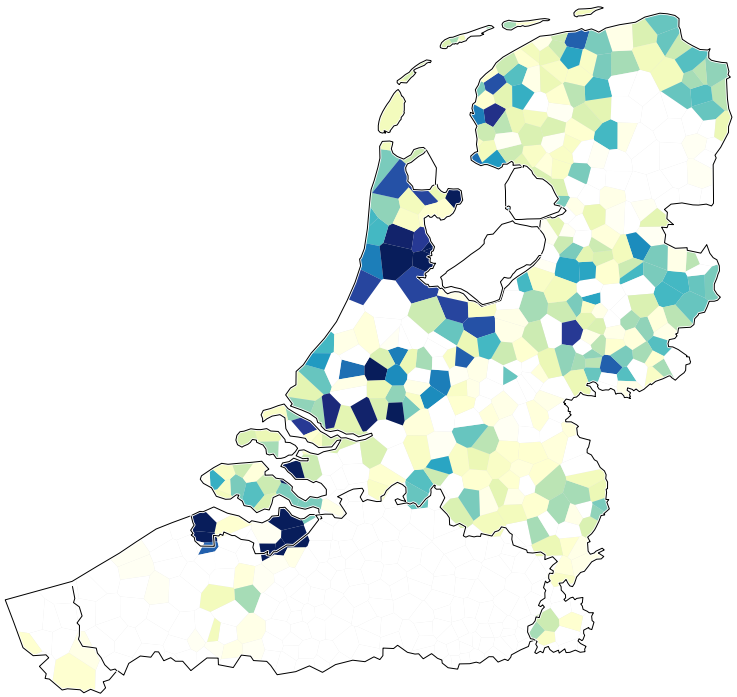
The data used in this walk-through consists of phonetic transcriptions of 562 items collected from 613 sites between 1979–1996 (Goeman and Taeldeman 1996).

Gabmap draws maps

- The typical work flow using gabmap includes visualizations on a map
- Gabmap accepts base map files in KML format

A look at the data

- Gabmap summarizes data in a few ways.
- It also visualizes distribution of certain features on the map.
- Example on the right shows the distribution of velarized lateral approximant [ɬ].
- It also supports regular expressions



How do we measure linguistic differences?

b	w	ʊ	t		r	
b		ɔ	t	ə	r	
	1	1		1		**3**

- For string (phonetic transcriptions) data, the minimum edit distance is used (customization of costs are also possible).
- For numeric data the Euclidean distance is used.
- For categorical data, either binary comparison, or a weighted similarity index can be used.

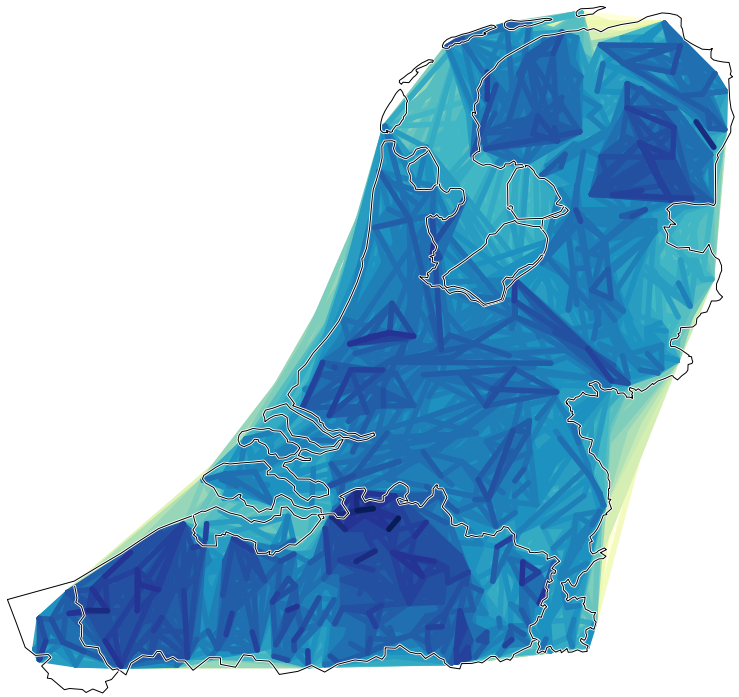
Aggregate linguistic differences

- The difference between two sites is calculated as the average difference between all items.
- This results in $\frac{N \times N}{2} - N$ distances for N sites ($187,884$ distances for our demonstration).

	Aals.	Baar.	Blan.	Coev.	...
Aalsmeer	0	0.26	0.30	0.24	...
Baardegem		0	0.21	0.28	...
Blankenberge			0	0.25	...
Coevorden				0	...
...	...	...	...	...	...

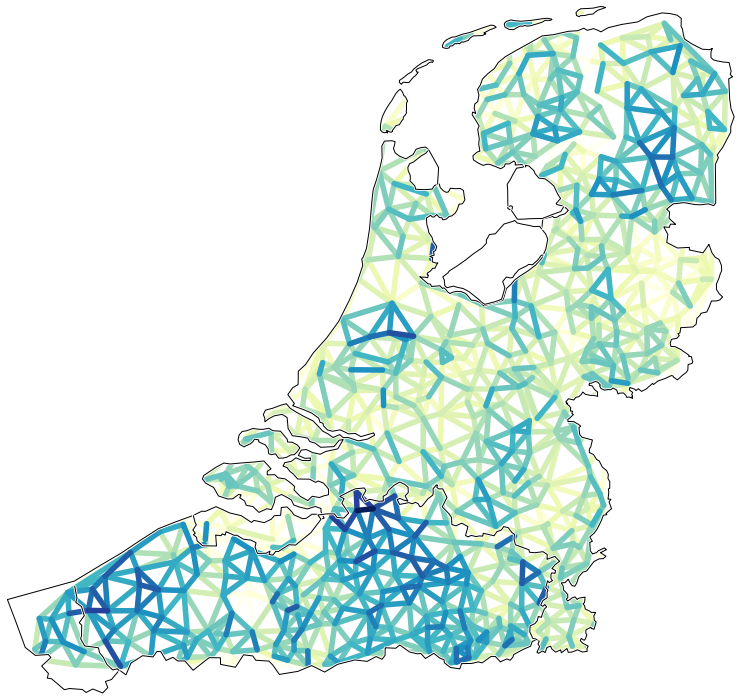
Aggregate linguistic differences

- The difference between two sites is calculated as the average difference between all items.
- This results in $\frac{N \times N}{2} - N$ distances for N sites ($187,884$ distances for our demonstration).

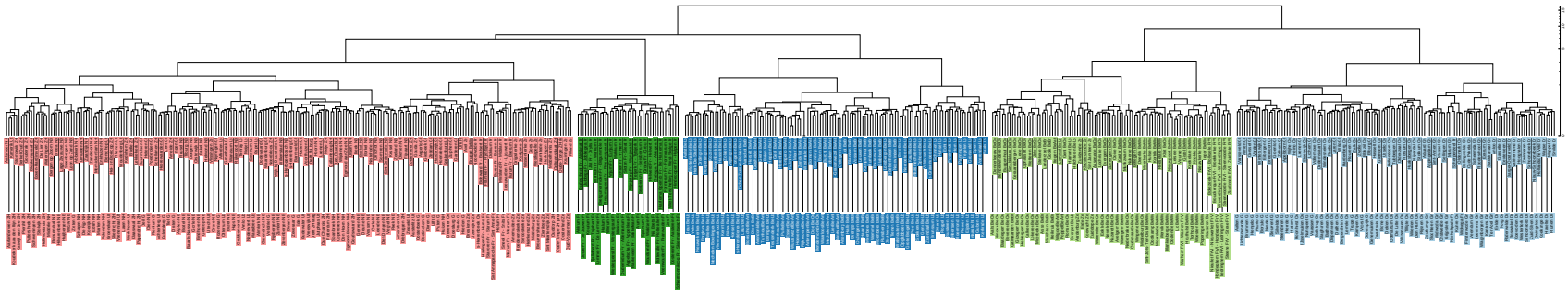


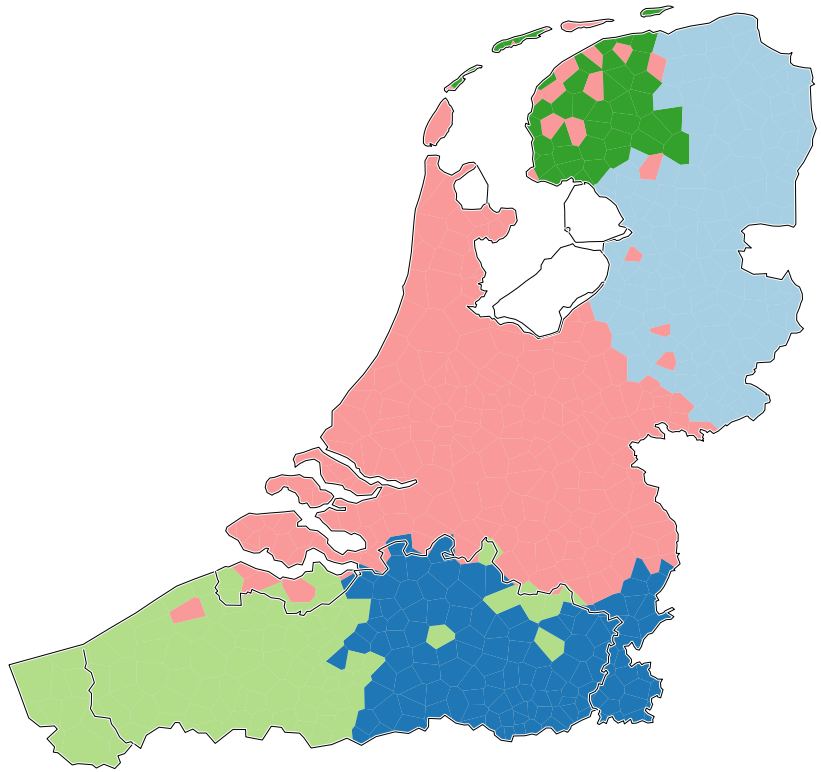
Aggregate linguistic differences

- The difference between two sites is calculated as the average difference between all items.
- This results in $\frac{N \times N}{2} - N$ distances for N sites ($187,884$ distances for our demonstration).



Clustering the sites



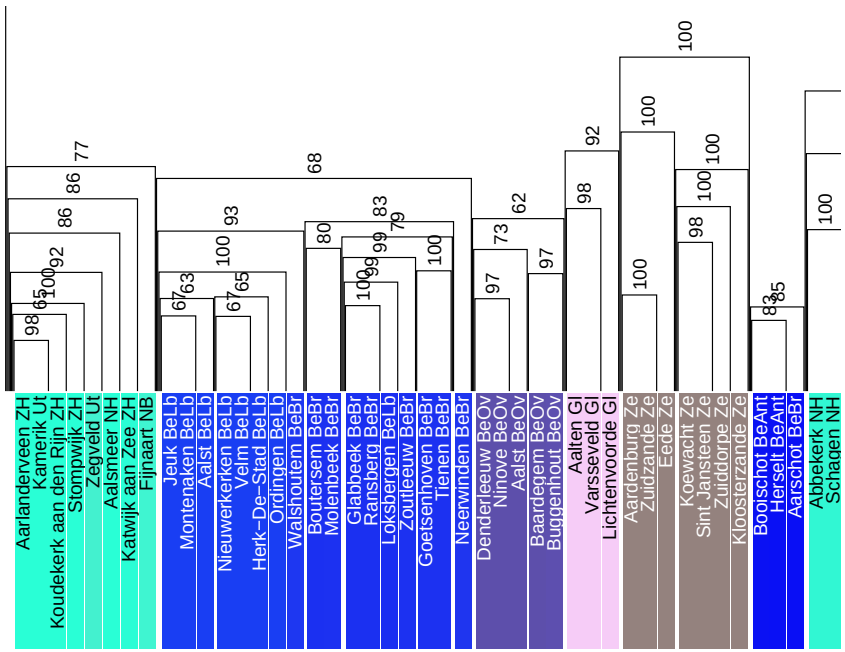


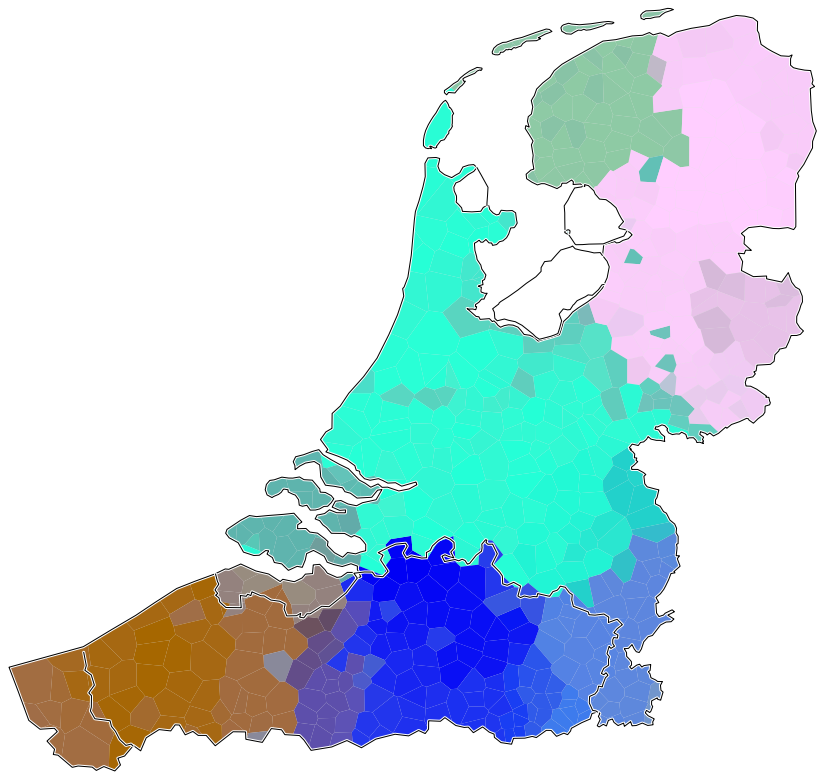
We often find expected results. However,

- Clustering is unstable
- Clustering does not necessarily indicate dialect boundaries

Probabilistic clustering

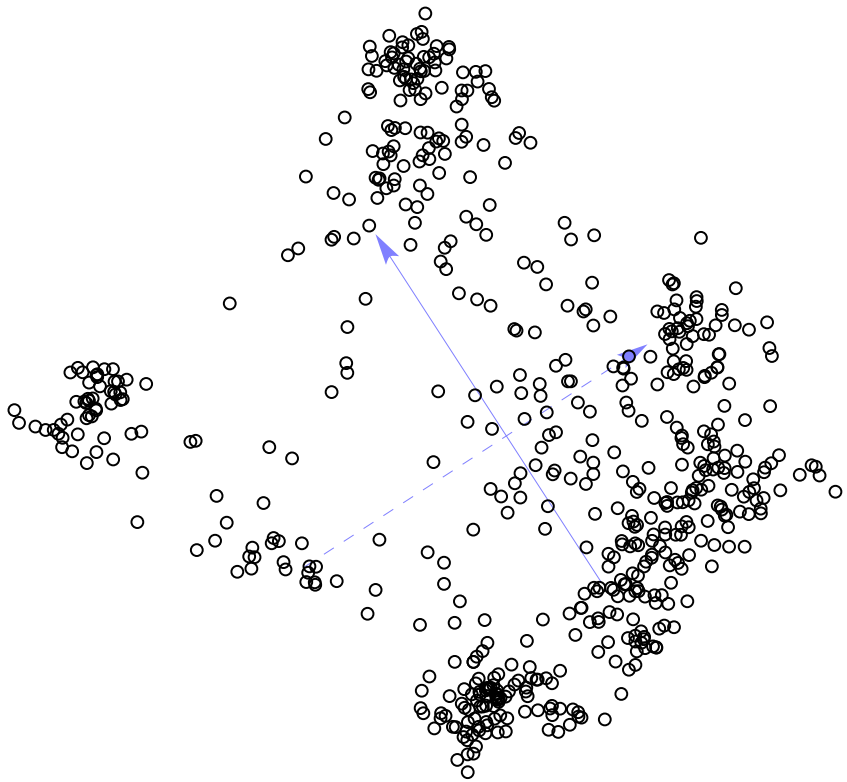
Probabilistic dendrograms can be used for checking the stability of the clusters.



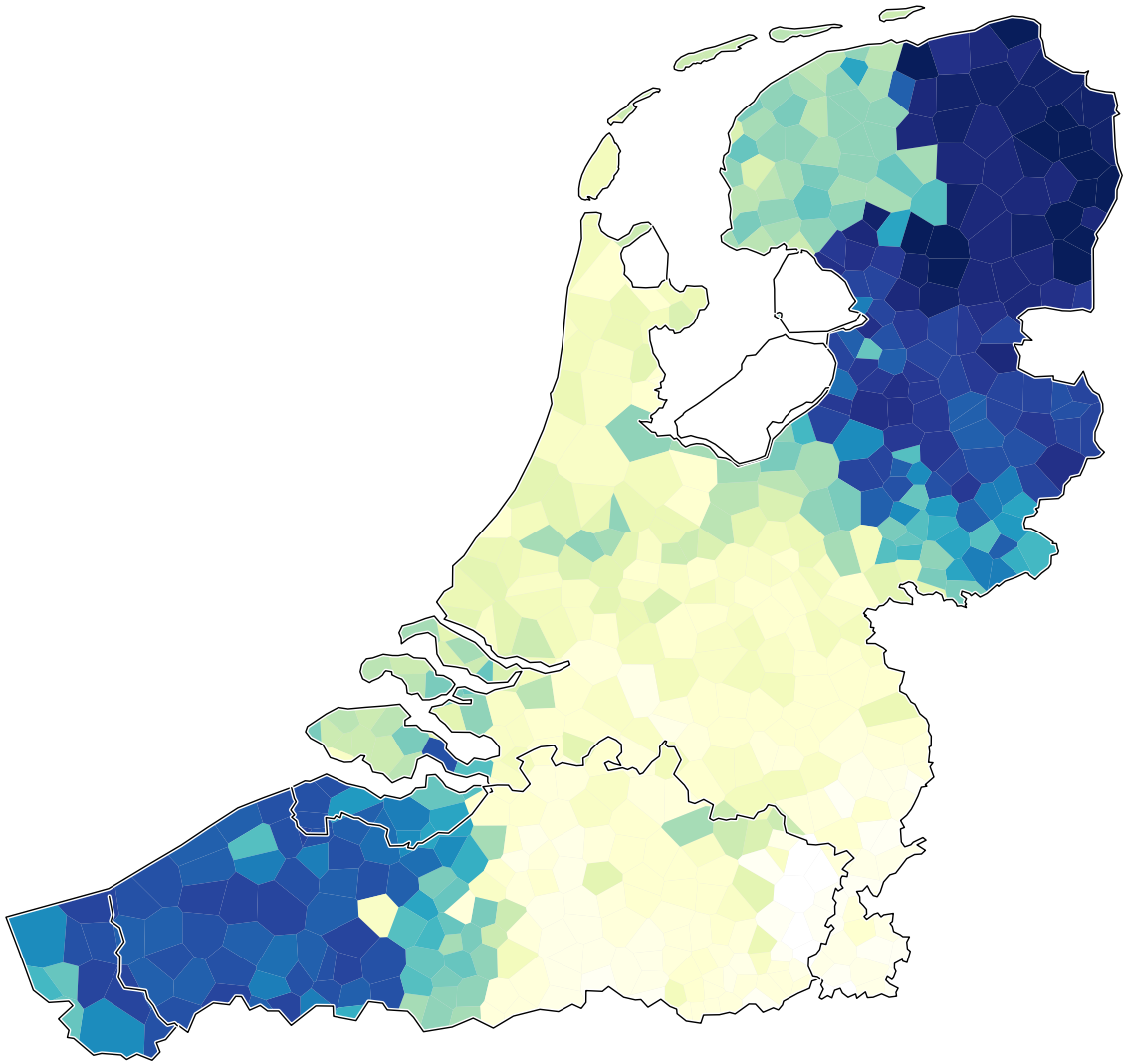


Multidimensional scaling (MDS)

- MDS is a technique to reduce the dimensionality of data while keeping the distances of the objects as close to the original as possible.
- MDS into 3 dimension explains 80–90% of the variation in most dialect data sets.
- MDS provides a good way to visualize the dialect continuum (or potential borders).

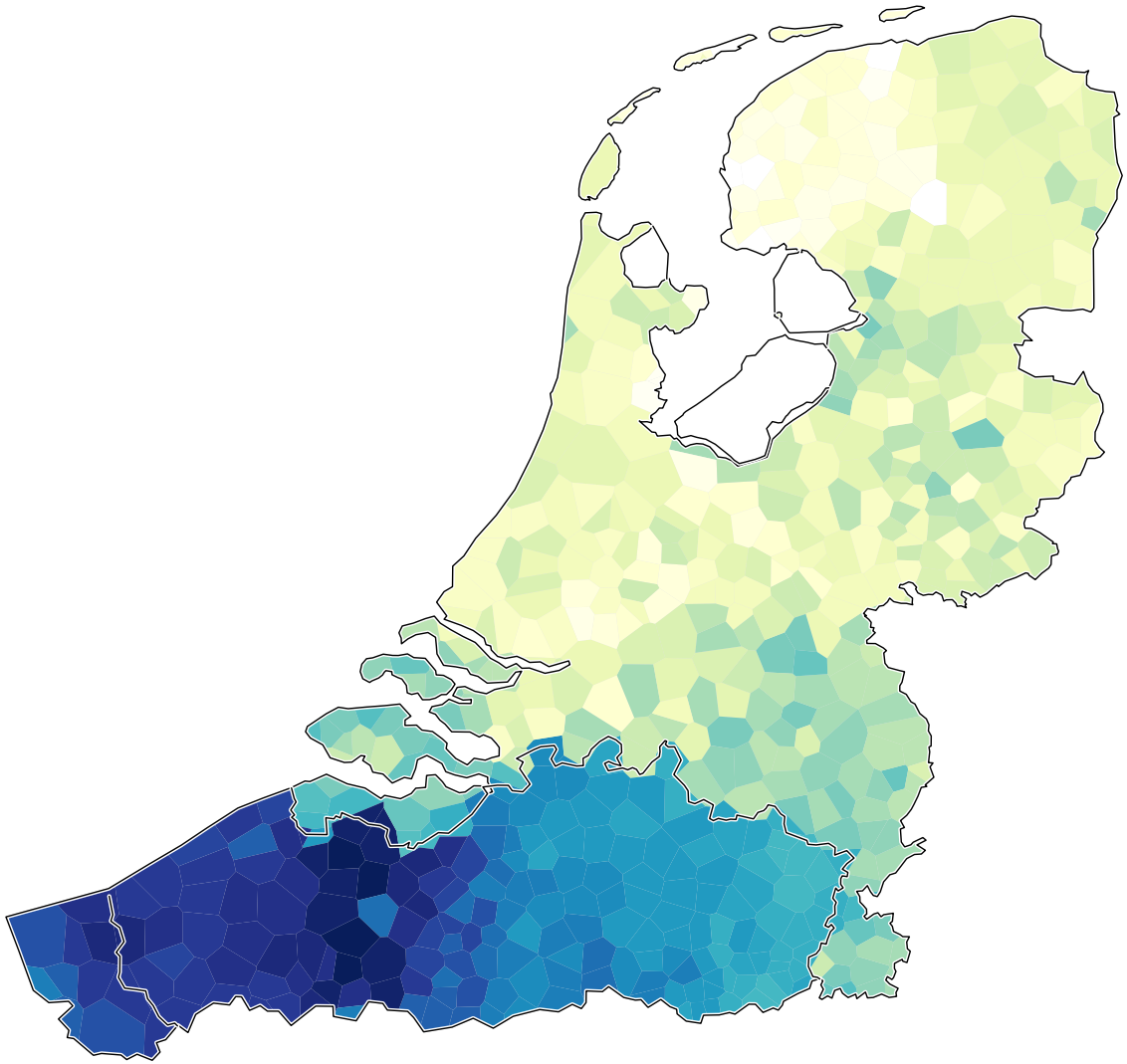


MDS maps - the first dimension



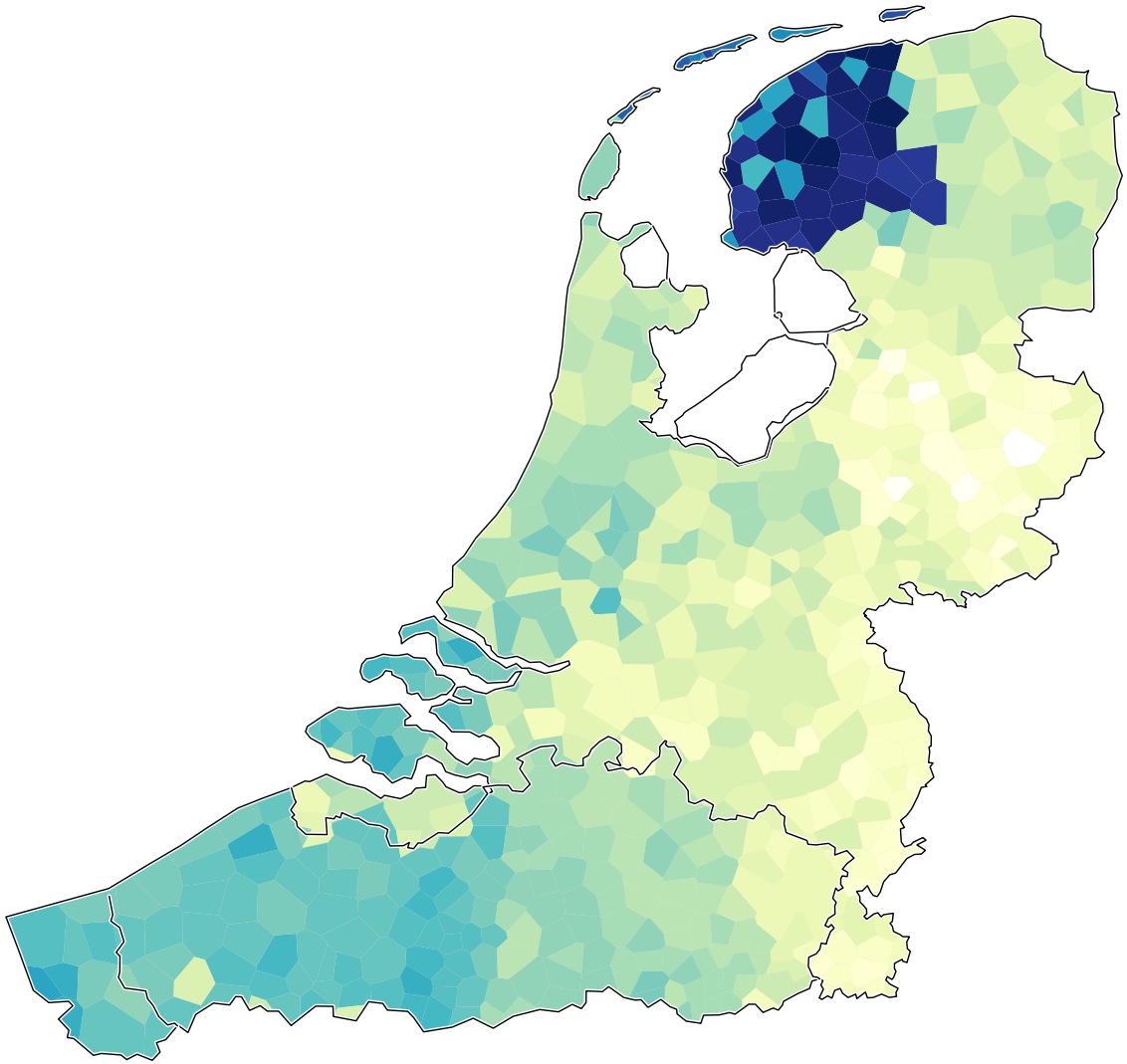
$r = 0.70$

MDS maps - the second dimension



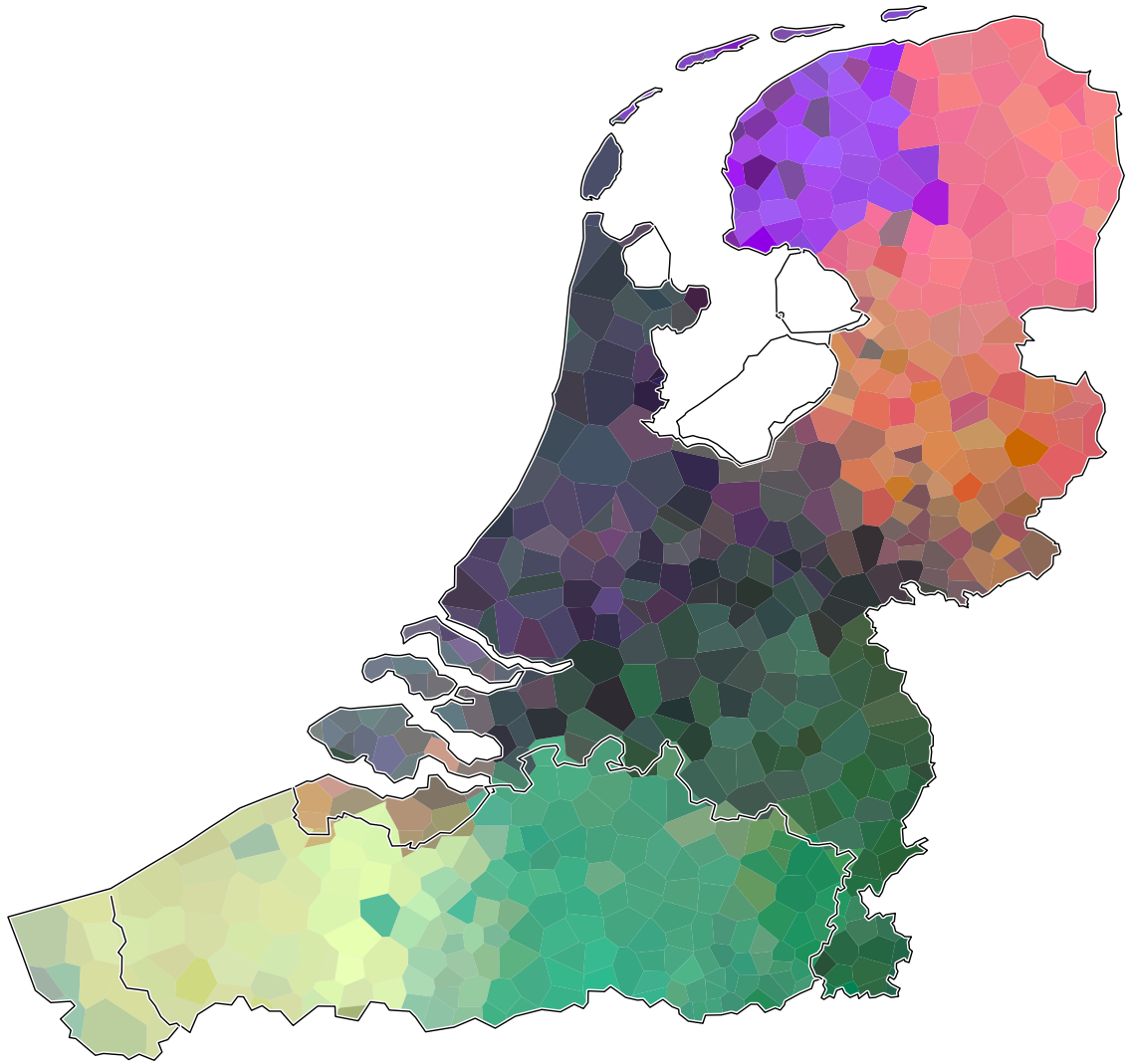
$r = 0.48$

MDS maps - the third dimension



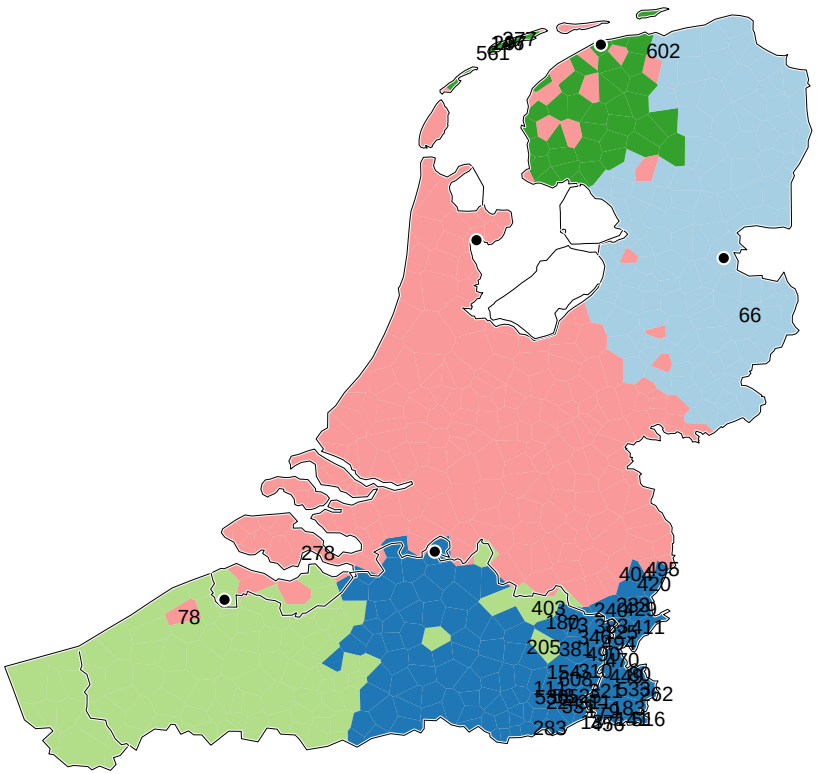
$r = 0.37$

MDS maps - the first three dimension



$r = 0.89$

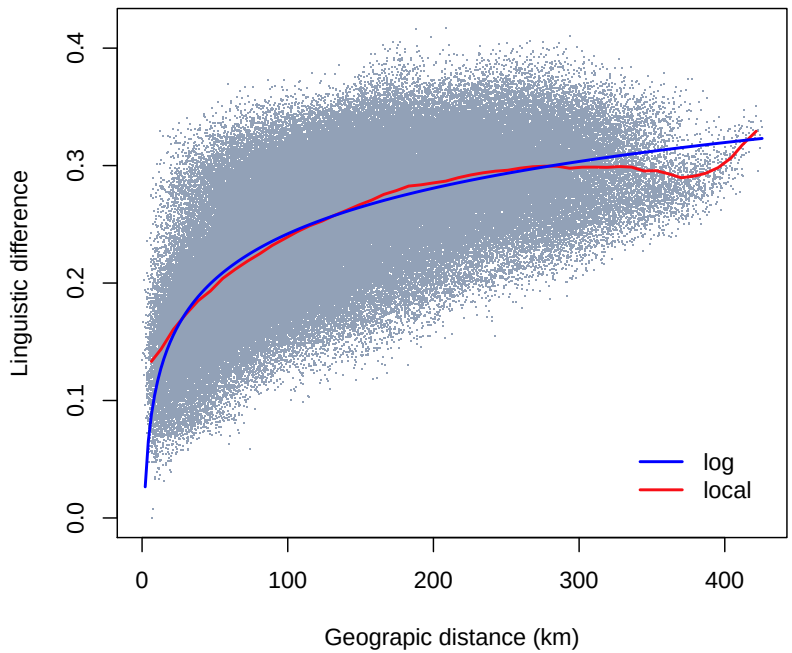
Using MDS for cluster validation



$$r = 0.83$$

Linguistic difference – geographical distance

- The exact relation between the linguistic differences and the geographical differences has been an interest in the study of linguistic change.
- Gabmap also allows an analysis for this purpose.



Finding shibboleths

> A **shibboleth** (/ˈʃɪbəlɛθ/ or /ˈʃɪbələθ/) is a word or custom whose variations in pronunciation or style can be used to differentiate members of ingroups from those of outgroups.
>
> — from Wikipedia

Finding shibboleths

> A **shibboleth** (/ˈʃɪbəlɛθ/ or /ˈʃɪbələθ/) is a word or custom whose variations in pronunciation or style can be used to differentiate members of ingroups from those of outgroups.
>
> — from Wikipedia

- We aim to find items (words) whose features (pronunciations) are characteristic for a dialect group.
- The item should be distinctive: the features associated with the item within the group should be different than the ones outside the group.
- The item should be representative: the feature should be uniform within the group of interest.

An example

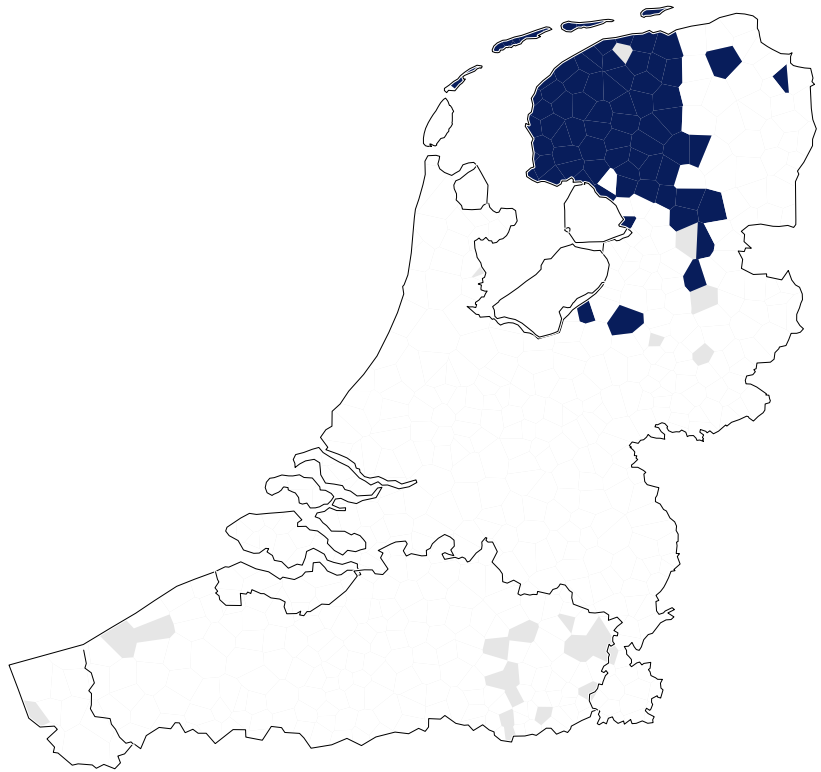
Top five 'shibboleths' for the Frisian dialect area.

Item	repr.	dist.	score
vinden	2.11	0.07	2.19
zoet	0.87	1.11	1.99
knieen	0.87	1.07	1.94
draden	1.34	0.52	1.87
bladeren	1.65	0.10	1.76

Example – distribution of particular forms

Pronunciation distribution of the word ‘vinden’ in Friesland (about 30 other forms are observed outside the Frisian area).

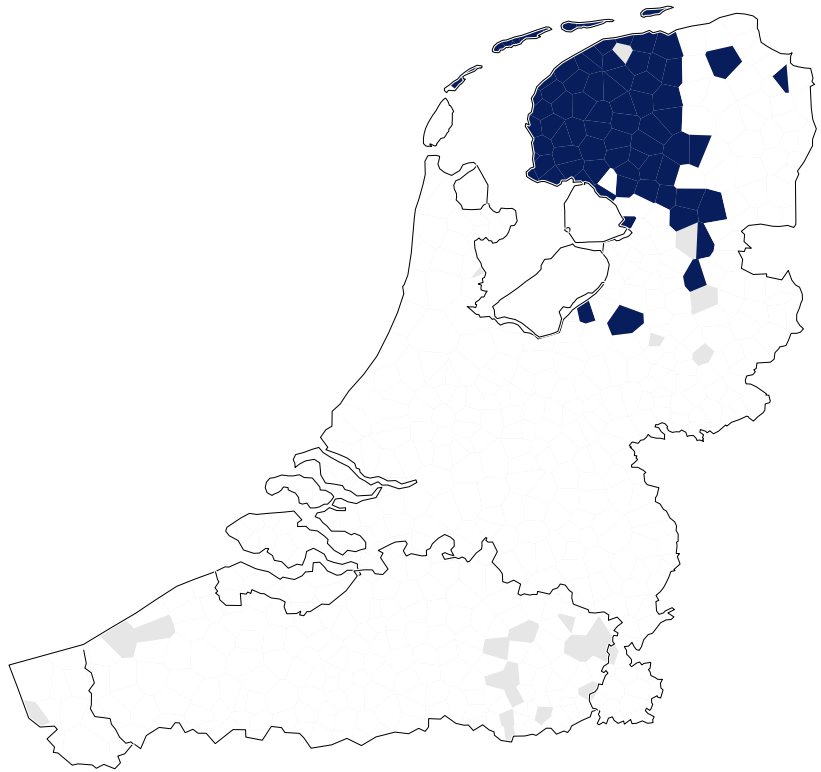
Pronunciation	in	out
finə	52	0
finn	1	0
fɪndn	1	2
fin	2	17



Example – distribution of particular forms

Pronunciation distribution of the word ‘vinden’ in Friesland (about 30 other forms are observed outside the Frisian area).

Pronunciation	in	out
finə	52	0
finn	1	0
fɪndn	1	2
fin	2	17



Further analysis shows that the combination of (1) dropping final [n] and initial [f] in words such as ‘vinden’ is characteristic to this area.

Gabmap – the application

- An easy-to-use web-based application: no computational expertise is needed
- Built on earlier tools, primarily RuG/L04 with some loss of flexibility, but also with many additions
- A web service is available. Anyone can use it after a simple registration (`http://gabmap.nl`)
- Open source, one can install it on another server, or change it to fit a particular need (source code at `https://github.com/coltekin/Gabmap`)

Gabmap – the ingredients (the scary stuff)

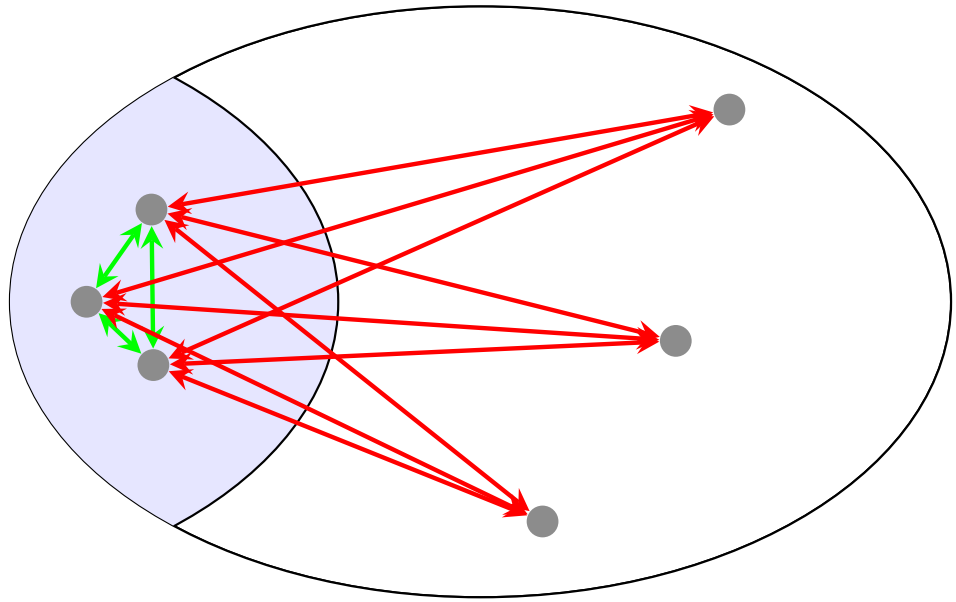
- Mainly written in Python 3, but also Python 2 in some cases
- Some use of JavaScript in the interface, and naturally, HTML/css for presentation of pages
- RuG/L04 for many tasks such as clustering (written in C/perl/flex...)
- R for some of the analyses
- UNIX shell scripts, makefiles
- Even some 'PostScript programs'

Gabmap – the ingredients (the scary stuff)

- Mainly written in Python 3, but also Python 2 in some cases
- Some use of JavaScript in the interface, and naturally, HTML/css for presentation of pages
- RuG/L04 for many tasks such as clustering (written in C/perl/flex...)
- R for some of the analyses
- UNIX shell scripts, makefiles
- Even some 'PostScript programs'

A good case for a web-based service?

Representativeness and distinctiveness



- The smaller the average difference between sites in the group (green links), more representative the feature is.
- The larger the average difference between sites in the group and outside (red links), more distinctive the feature is.

Mathematical definition

Average within-group difference

$$\bar{d}_f^g = \frac{2}{|g|^2 - |g|} \sum_{s,s' \in g} d_f(s, s')$$

Average between-group difference

$$\bar{d}_f^{g'} = \frac{1}{|g|(|G| - |g|)} \sum_{s \in g, s' \notin g} d_f(s, s')$$

A combined and normalized measure

$$\underbrace{\frac{\bar{d}_f^{g'} - \bar{d}_f}{sd(d_f)}}_{\text{distinctiveness}} \quad \underbrace{- \quad \frac{\bar{d}_f^{g} - \bar{d}_f}{sd(d_f)}}_{\text{representativeness}}$$

Bibliography

Goeman, Antonie and Johan Taeldeman (1996). "Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten". In: *Taal en Tongval* 48, pp. 38–59.

Nerbonne, John (2013). "How much does geography influence language variation?" In: ed. by Peter Auer et al. Berlin: Mouton de Gruyter, pp. 220–36.

Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne (2012). "Detecting shibboleths". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, pp. 72–80.

Séguy, Jean (1971). "La relation entre la distance spatiale et la distance lexicale". In: *Revue de Linguistique Romane* 35.138, pp. 335–357.

Therese Leinonen, Çağrı Çöltekin and John Nerbonne (2015). "Using Gabmap". In: *Lingua* (to appear). Special issue on linguistic infrastructure.