Methods for aggregate corpus-based dialectometry

Natalia Levshina

Mainz, May 10 2016

What is dialectometry?

- Introduced by Séguy (1971)
- Examines the differences and similarities between geographical locations with regard to a large number of linguistic variables
  - Strong agreement between neighbouring locations suggests a dialect area.
  - Weak agreement between neighbouring locations suggests a dialectal boundary.

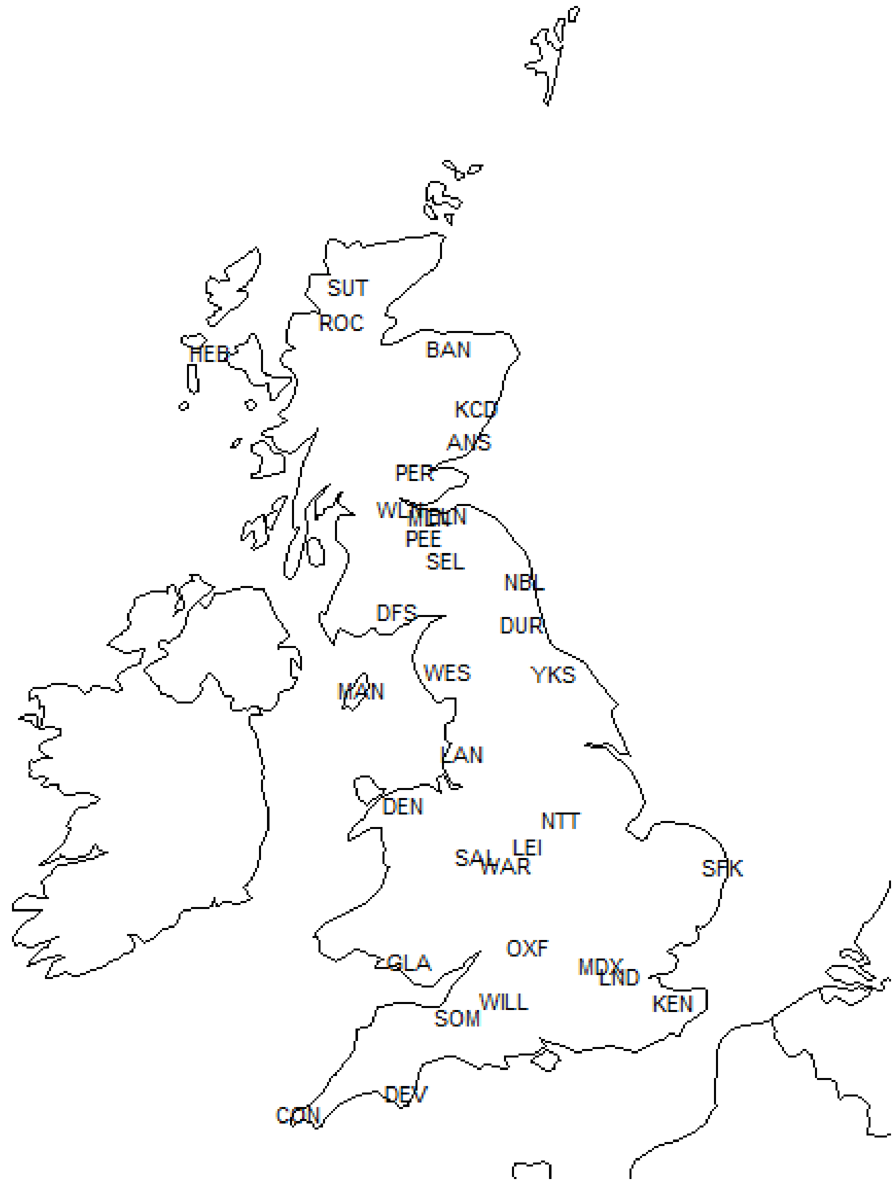
Main inspiration

Outline

1. Data: different approaches
    - Szmrecsanyi (2013) dataset of BrE dialects
    - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
    - Preliminary univariate analyses
    - Aggregate distance matrix
    - Exploration of dialect continua (MDS)
    - Identification of dialect areas (cluster analysis)

Szmrecsanyi (2013) data

- Available from
  https://sites.google.com/site/bszmrecsanyi/datasets
- The Freiburg Corpus of British dialects (FRED): interviews (interviewers' utterances excluded)
- 158 different locations in 34 counties (pre-1974) of Great Britain + the Isle of Man and the Hebrides
- 57 morphosyntactic features


SUT
ROC
HEB
BAN
KCD
ANS
PER
WLN
MDL
ELN
PEE
SEL
NBL
DFS
DUR
WES
YKS
MAN
LAN
DEN
NTT
LEI
SAL
WAR
SFK
OXF
GLA
MDX
LND
WILL
KEN
SOM
DEV
CON

Groups of features with examples

A. Pronouns and determiners.

- [1] non-standard reflexives, e.g. *They didn't do it theirself.*

B. The noun phrase

- [9] the '*s*-genitive, e.g. John'*s* book

C. Primary verbs

- [15] the primary verb TO HAVE, e.g. *The time has passed.*

D. Tense and aspect

- [20] used to as a marker of habitual past, e.g. *He used to go around killing pigs.*

Groups of features (cont.)

E. Modality

F. Verb morphology

G. Negation

H. Agreement

I. Relativization

J. Complementation

K. Word order and discourse phenomena

Selection of features

- Dialectological, variational and corpus-linguistic literature
- Non-standard and standard features
- Alternation variables, e.g. standard and non-standard reflexives (e.g. *theirself – themselves*), and non-alternating forms (e.g. preposition stranding, *The house I lived in*)

Frequency matrix

- The extracted instances of variables are summarized for each county (34 in total)
- Normalized (per 10K words in the corpus)
- Log_{10}-transformed (to reduce the effect of large frequency differences and emphasize small frequency differences)
- N x p matrix (N = number of objects, i.e. 34 counties, p is the number of features, 57)

A fragment of the data set

County code	County	avg_longitude	avg_latitude	a1_nonstandard_refl	a2_standard_refl	a3_thee_thou_thine	a4_ye
ANS	Angus	-2.6269	56.6594	0	0.845098	0	0.6020
BAN	Banffshire	-2.9495	57.5431	0.3010299	0.602059	0.3010299	-1
CON	Cornwall	-5.5021	50.1754	0	0.698970	-1	-1
DEN	Denbighshire	-3.7429	53.1463	-1	0.698970	-1	-1
DEV	Devon	-3.681	50.3777	0	0.778151	-1	-1
DFS	Dumfriesshire	-3.8395	55.0028	-1	0	0	-1

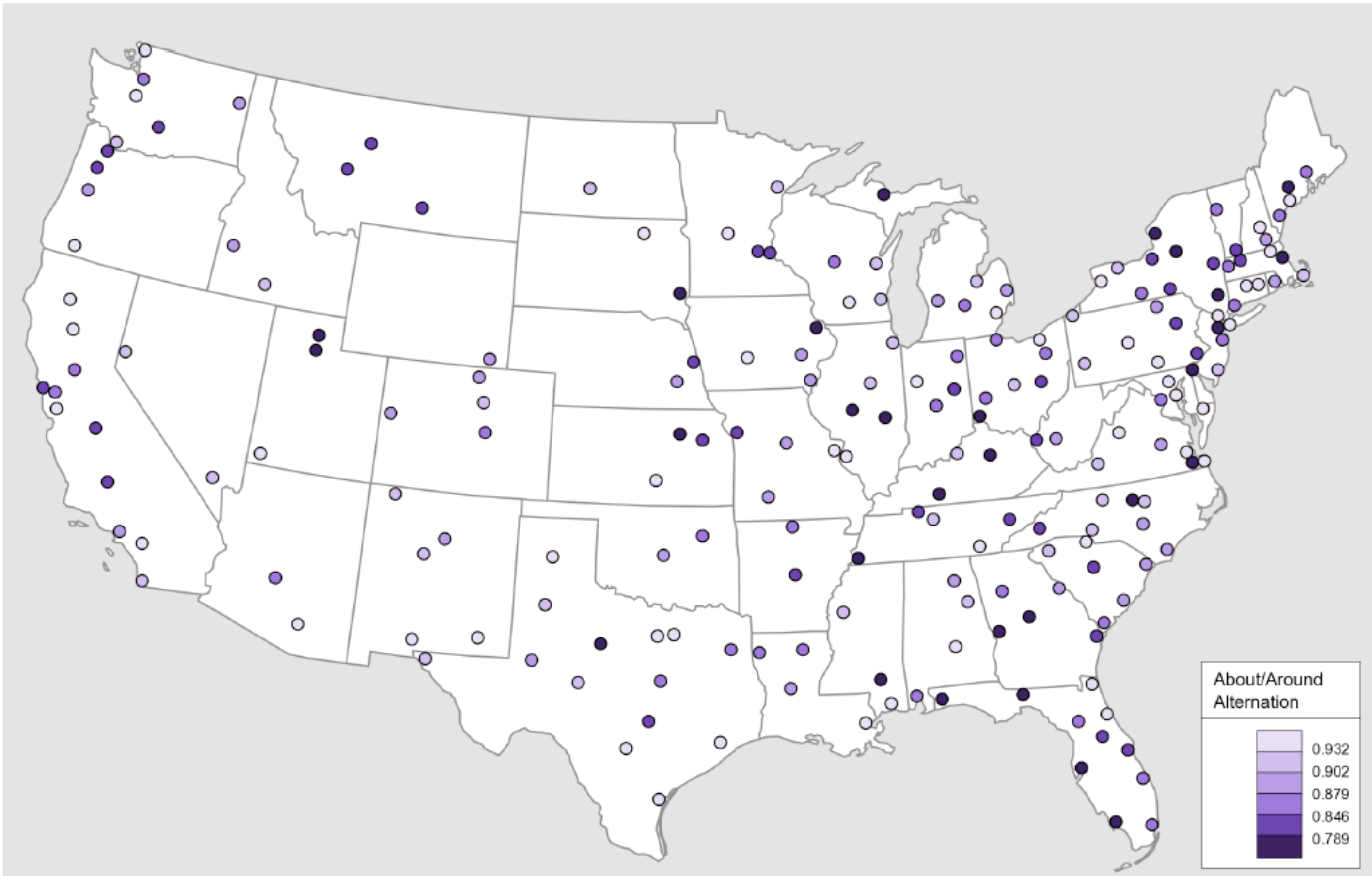
Outline

1. Data: different approaches
    - Szmrecsanyi (2013) dataset of BrE dialects
    - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
    - Preliminary univariate analyses
    - Aggregate distance matrix
    - Exploration of dialect continua (MDS)
    - Identification of dialect areas (cluster analysis)

Grieve (2009) data set

- Regional variation in written American English
- 206 cities from across the U.S.
- Letters to editors in local newspapers
- 40 lexical alternation variables only, e.g. though/although, whilst/while, about/around + Num
- The values in the matrix are the proportions of one variant (e.g. though) divided by the sum frequency of both variants (e.g. though + although) (from 0 to 1)
- The approach is more Labov-style.

Figure 3 About/Around Alternation Raw Values



From Grieve, Speelman & Geeraerts (2011)

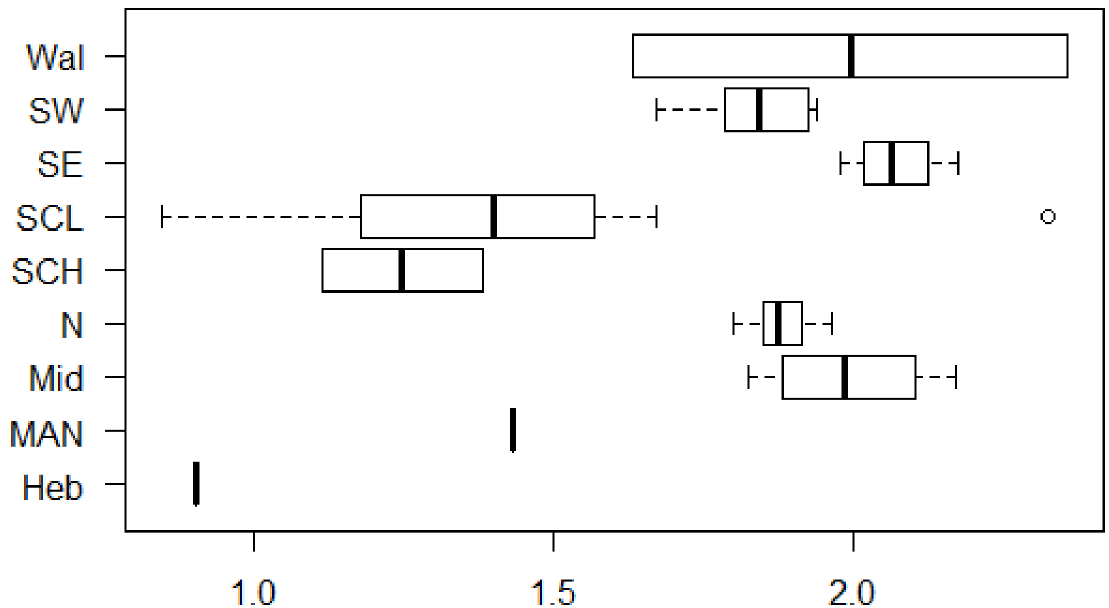
Outline

1. Data: different approaches
   - Szmrecsanyi (2013) dataset of BrE dialects
   - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
   - Preliminary analyses
   - Aggregate distance matrix
   - Exploration of dialect continua (MDS)
   - Identification of dialect areas (cluster analysis)

How reliable is the frequency matrix?

- Cronbach's α: often used as a measure of reliability of test scores in psychology.
- A high score means that the variables are sufficiently intercorrelated for performing further analyses.
- Conventional minimum: $\alpha = 0.7$
- Here: $\alpha = 0.77$

Boxplot by areas: Var20 *used to*



Correlation with geographic distances

- Step 1. For each variable separately, compute pairwise linguistic distances by using the Euclidean metric. In this case (1 variable), the distances are absolute frequency differences between the points.

	Chapman_code	county	a20_USED_TO_habitual_past
1	ANS	Angus	1.477121
2	BAN	Banffshire	0.845098
3	CON	Cornwall	1.845098
4	DEN	Denbighshire	2.359835
5	DEV	Devon	1.924279
6	DFS	Dumfriesshire	2.328380
7	DUR	Durham	1.913814

Distance matrix (Var20)

	ANS	BAN	CON	DEN	DEV	DFS
BAN	0.632					
CON	0.368	1.000				
DEN	0.883	1.515	0.515			
DEV	0.447	1.079	0.079	0.436		
DFS	0.851	1.483	0.483	0.031	0.404	
DUR	0.437	1.069	0.069	0.446	0.010	0.415

Step 2. Compute geographical distances from the coordinates

	ANS	BAN	CON	DEN	DEV	DFS
BAN	62					
CON	464	520				
DEN	247	306	219			
DEV	437	497	82	192		
DFS	124	179	341	128	320	
DUR	128	190	363	146	323	85

(the distances are in rounded miles, taking into account the Earth curvature)

Step 3. Correlation between linguistic and geographic distances

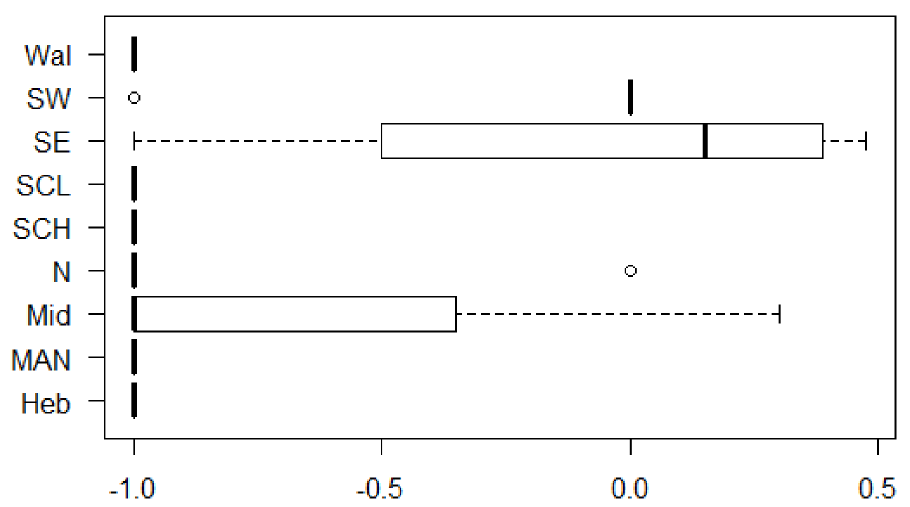
- Mantel's permutation test for distance matrices:

  $r = 0.43$, $p = 0.001$

- Therefore, the closer are two locations, the more similar frequencies of the feature they have.

Exercise. Task 1

- Var32, the use of *ain't*, is distributed across the regions as shown below. What does the boxplot say?



Exercise. Task 2

- What is the linguistic distance between Kent and Suffolk with regard to *ain't*?

	Chapman_code	apriori_area	a32_aint
13	KEN	SE	0.4771213
27	SFK	SE	0.3010300

- Is there a significant correlation between the linguistic and geographic distances?
  - Mantel r = 0.3, p = 0.001

Outline

1. Data: different approaches
   - Szmrecsanyi (2013) dataset of BrE dialects
   - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
   - Preliminary univariate analyses
   - Aggregate distance matrix
   - Exploration of dialect continua (MDS)
   - Identification of dialect areas (cluster analysis)

Euclidean distance

- Computed for every pair of locations
- The distance represents:
  - the square root…
  - of the sum…
  - of the squared differences between two locations for every variable

An example with two variables

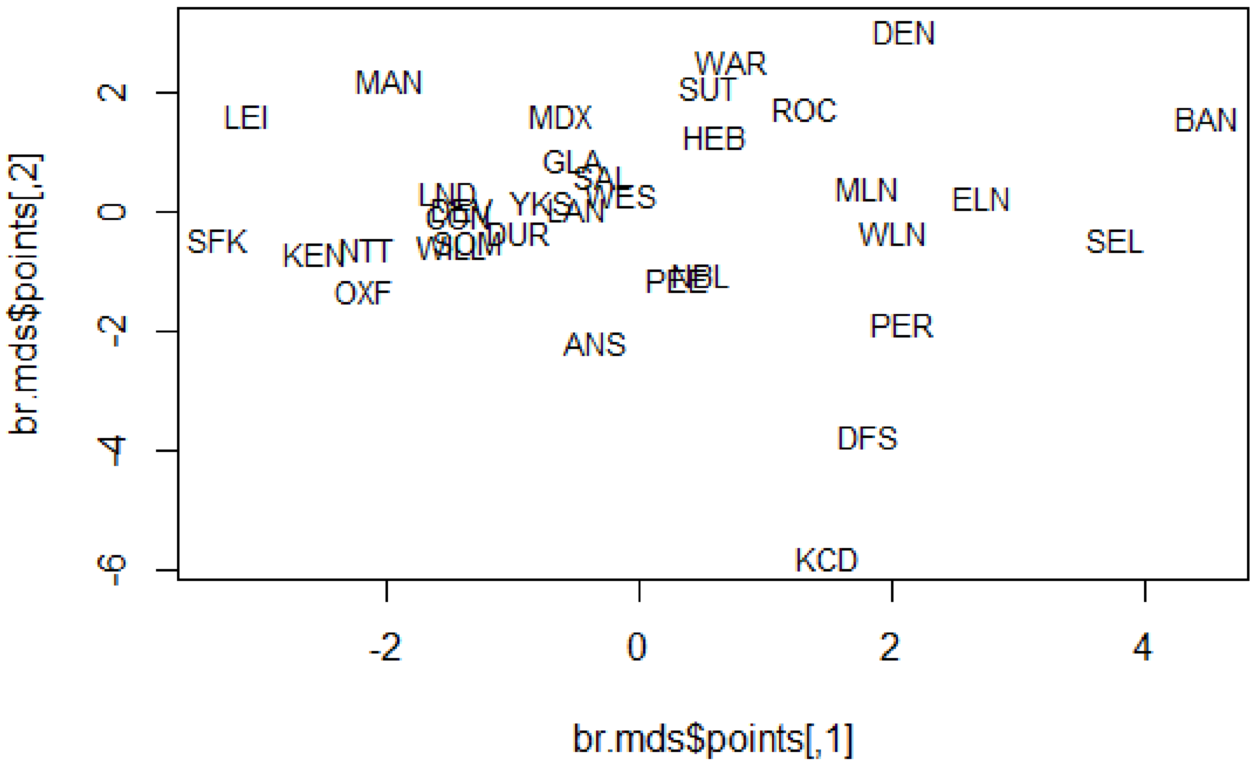
code	county	a20_USED_TO_hab.past	a32_aint
KEN	Kent	2.176091	0.4771213
SFK	Suffolk	2.075547	0.3010300

```
> sqrt((2.176091 - 2.075547)^2 + (0.4771213 - 0.3010300)^2)
[1] 0.2027739
```

Outline

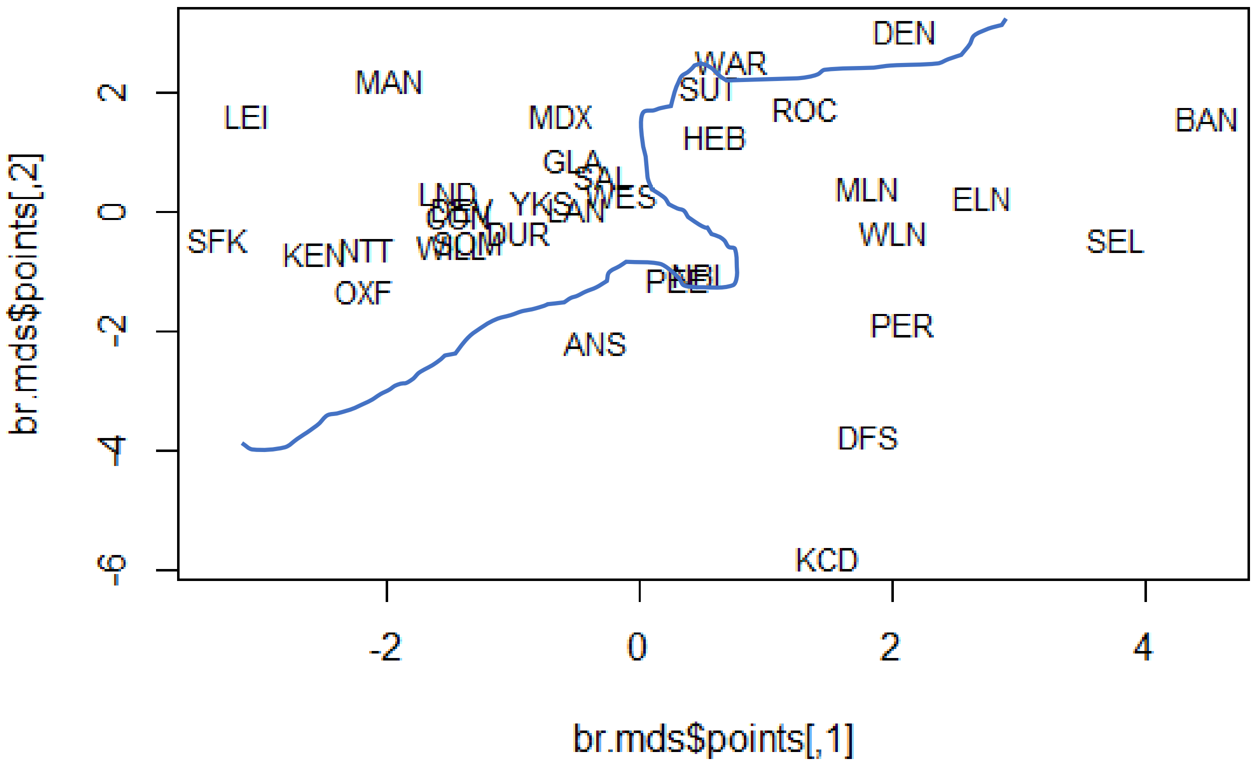
1. Data: different approaches
   - Szmrecsanyi (2013) dataset of BrE dialects
   - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
   - Preliminary univariate analyses
   - Aggregate distance matrix
   - Exploration of dialect continua (MDS)
   - Identification of dialect areas (cluster analysis)

MDS (non-metric, 2 dimensions)



Stress = 0.18

[England + Wales] vs. [Scotland]



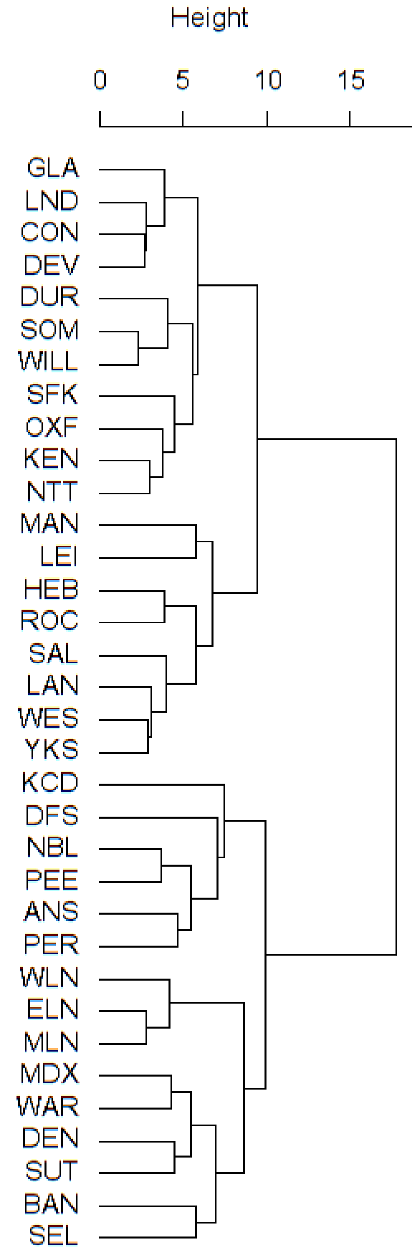
Stress = 0.18

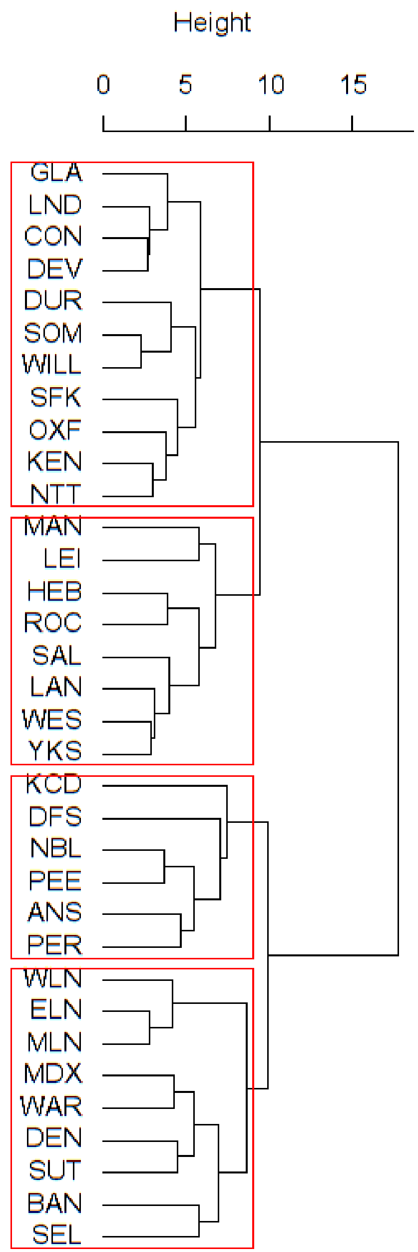
Outline

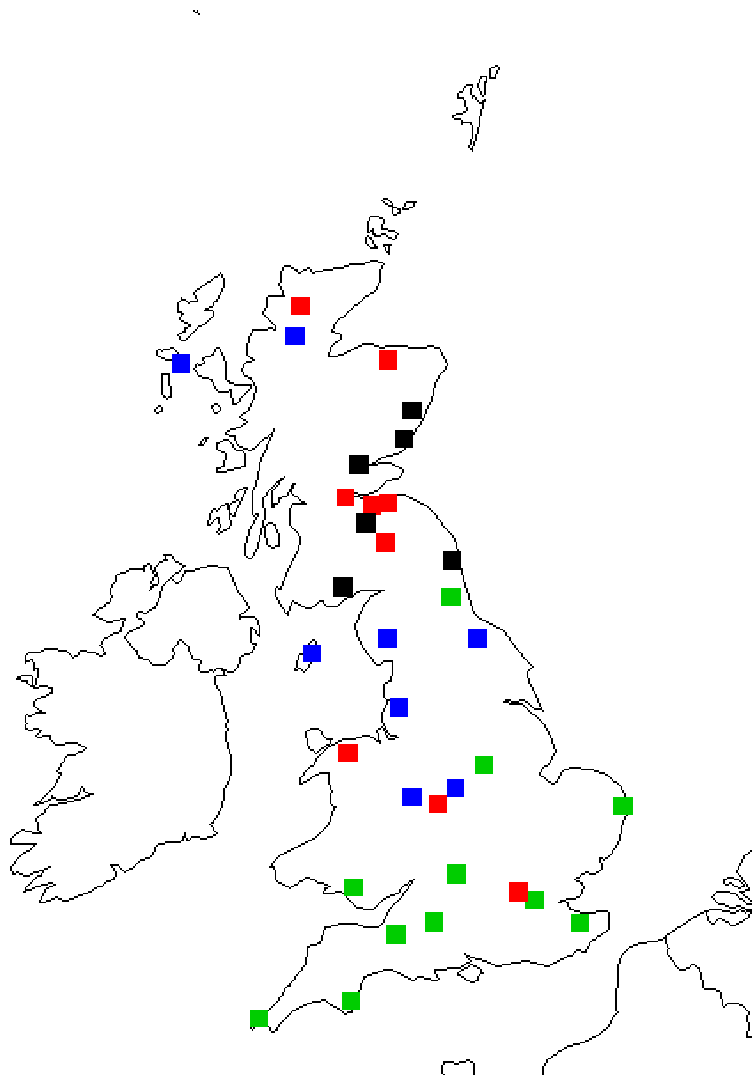
1. Data: different approaches
   - Szmrecsanyi (2013) dataset of BrE dialects
   - Grieve (2009) dataset of AmE dialects
2. Statistical analyses
   - Preliminary univariate analyses
   - Aggregate distance matrix
   - Exploration of dialect continua (MDS)
   - Identification of dialect areas (cluster analysis)

Hierarchical cluster analysis (Ward method)

- requires a distance matrix
- first merges two closest points, then finds the next closest, and repeats until all points are merged
- different methods of choosing the nearest point/cluster to another cluster (complete, average, single, Ward’s, etc.)

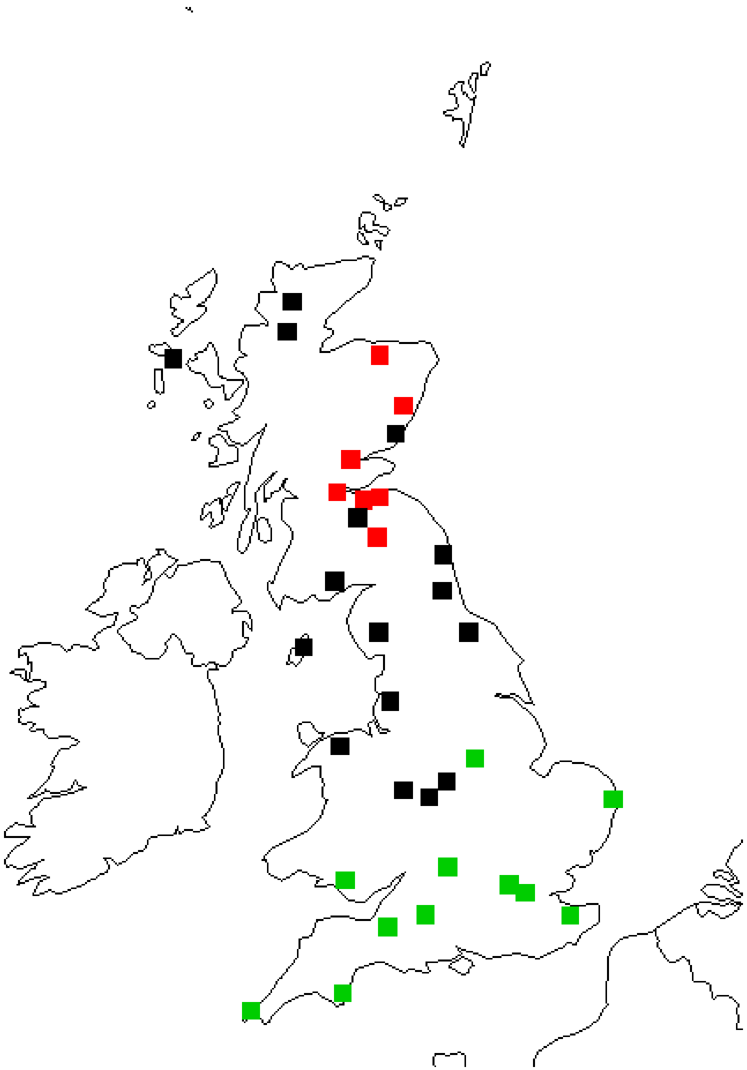

Height
0
5
10
15
GLA
LND
CON
DEV
DUR
SOM
WILL
SFK
OXF
KEN
NTT
MAN
LEI
HEB
ROC
SAL
LAN
WES
YKS
KCD
DFS
NBL
PEE
ANS
PER
WLN
ELN
MLN
MDX
WAR
DEN
SUT
BAN
SEL


Height
0
5
10
15
GLA
LND
CON
DEV
DUR
SOM
WILL
SFK
OXF
KEN
NTT
MAN
LEI
HEB
ROC
SAL
LAN
WES
YKS
KCD
DFS
NBL
PEE
ANS
PER
WLN
ELN
MLN
MDX
WAR
DEN
SUT
BAN
SEL



Divisive non-hierarchical clustering

- The algorithm divides the points into *n* clusters such that the distances between the clusters are maximized and the distances between the cluster members are minimized.
- PAM: Partitioning Around Medoids
  - can be performed either on a distance matrix, or on the frequency matrix
  - robust to outliers



Additional tools

- Groningen software for making fancy maps:
  - http://www.let.rug.nl/~kleiweg/L04/

References

- Grieve, J.. 2009. A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English. Ph.D. Dissertation. Northern Arizona University.
- Grieve, J., Speelman, D. & Geeraerts, D. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 193-221.
- Séguy, J. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335–357.
- Szmrecsanyi, B. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.