Finding transition zones in dialect data

Erik Tjong Kim Sang
Meertens Institute

18 September 2014

Postdoc task in the project Maps and Grammar

- Detect regions based on existing linguistic database with location information
- Visualize complex patterns present in linguistic data in geographic space

Case study: work with syntactic data and compare results with related earlier work done with pronunciation data

SAND: Syntactische Atlas van Nederlandse Dialecten

(Syntactic Atlas of Dutch Dialects)

- More than 700 syntactic variables/values measured in 267 locations with some dialect of Dutch
- Data are available on printed and digital maps, in text files and in database format
- All interviews on which the linguistic data was based, are available in audio format with transcriptions

See: `http://www.meertens.knaw.nl/sand`

Finding regions in geographic data

Step 1: convert data to vectors:

(Amsterdam,no,yes,no,no,no,yes,...)

Step 2: run a cluster algorithm on the data:

The vector of Amsterdam looks like the one of Zaandam so we will put them in the same region.

But the Amsterdam vector is quite different from the one of of Maastricht so they belong to different regions.

Step 3: visualize the regions

Interpreting SAND data

We need binary information for building models:

yes the syntactic variable occurs in this location
no the syntactic variable does not occur in this location

However, we also found:

this syntactic variable has value **A** at this location
this syntactic variable has value **B** at this location
this syntactic variable can take both values **A** and **B** at this location
the possible values of this syntactic value at this location are **unknown**

How do we deal with multivalued variables?

Multivalued data does not have explicit negative information

Previous work [Spruit 2008] dealt with this by making the following assumption:

> When a value of a syntactic variable has not been observed in a certain location then it does not occur at that location

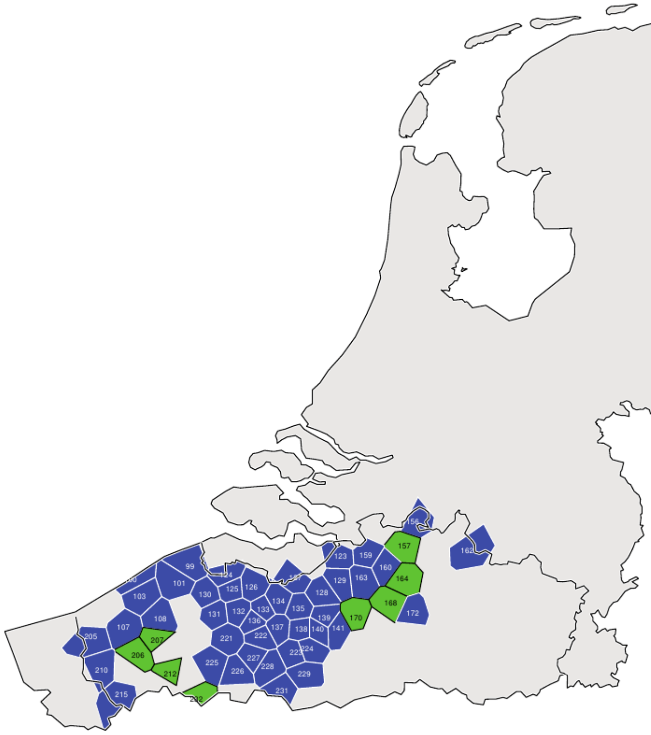
We do not want to make this assumption so we are only using the part of the SAND with binary values (yes vs no)

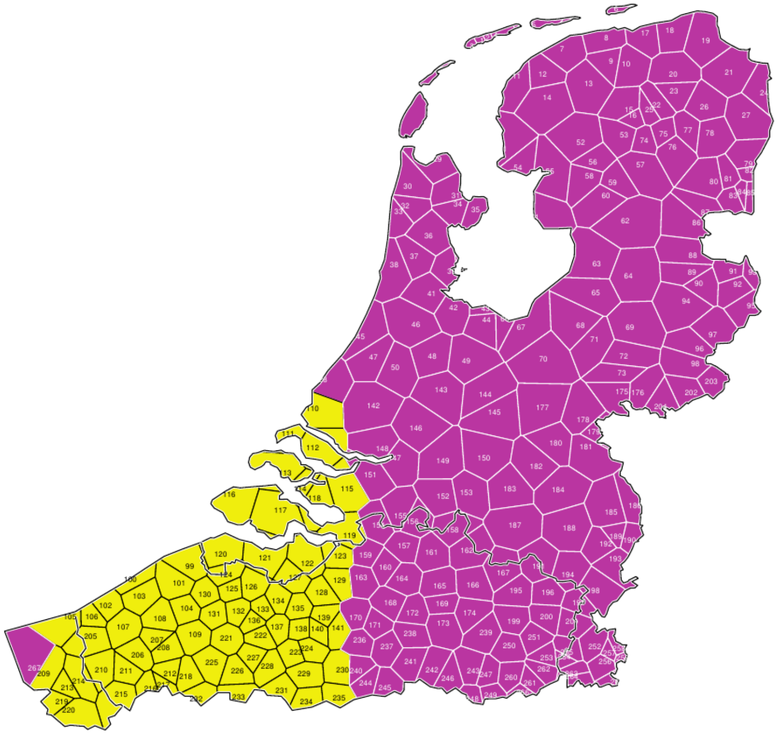
Which parts of the SAND are we using?

Map data values type	SAND1	SAND2
Only positive	(43b 50b) 64b 77b 79a (93a) 93b (94a) 95a	32b 41b 42a 42b 48b 49a 50b 51b (53a) 55a 57a
Positive and negative	(19a 22a 24a 26a 28a 30a 32a 87a 87b)	28a 28b 29a 29b 29c 30a (44a) 45b 46a 46b 46c
Multivalued and negative	39a 39b 40a 41b 43a 44b 45a 45b 46b 47b 48a 48b	19b 20b 45a 61b

Numbers represent pages; gray page numbers have not been used

How do we deal with missing values?





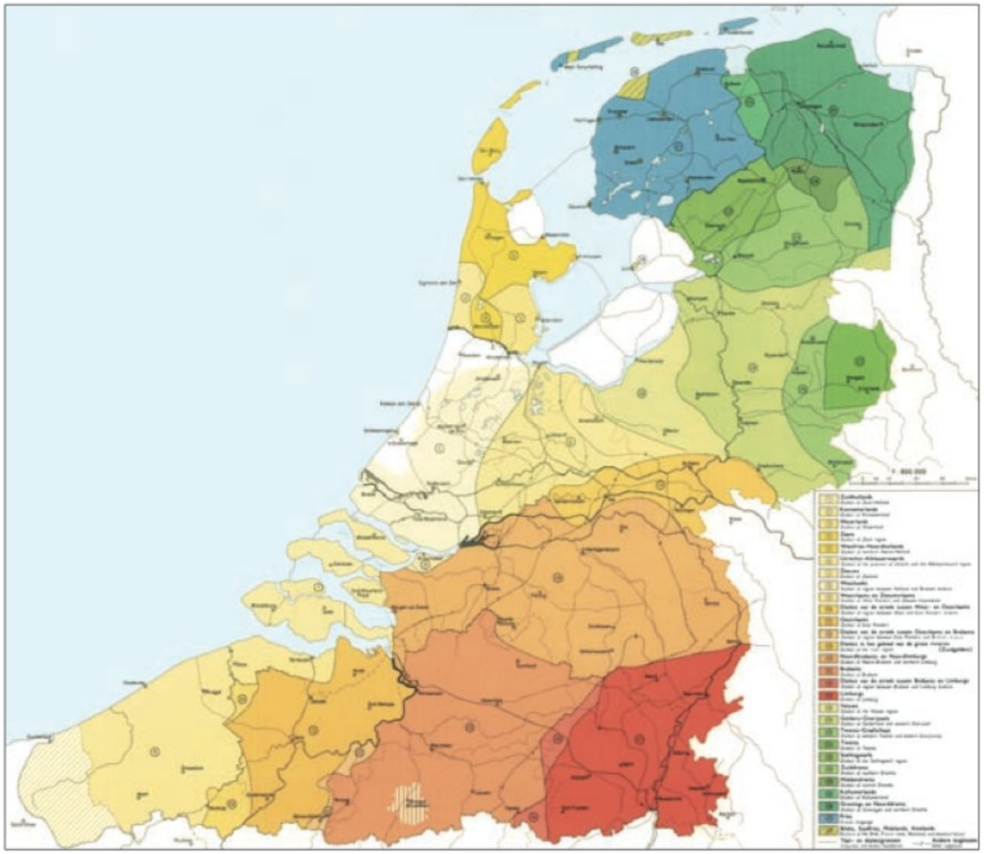
An alternative for clustering

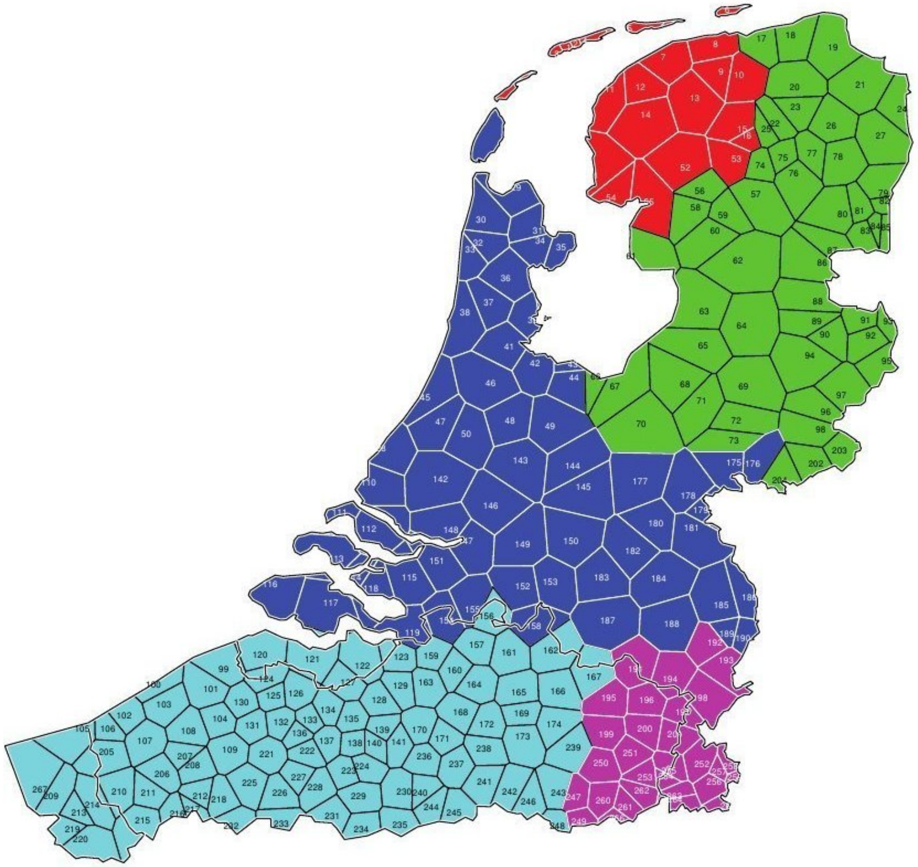
We are not using clustering techniques for processing the SAND data

- because of the large number of missing values
- because clustering results are difficult to interpret for a someone that is not a domain expert

Instead we opted for using supervised machine learning

We needed gold standard dialect regions and for this purpose we used a simplification of the 1969 Daan Blok Dutch dialect division based on pronunciation





Supervised Machine Learning: Naive Bayes

We used a basic machine learning technique for building a model of the relation between the syntactic variables and the gold standard regions: Naive Bayes

It counts how often variable values cooccur with regions and assigns a weight to the relation between the two based on how often they occur together and how often they do not

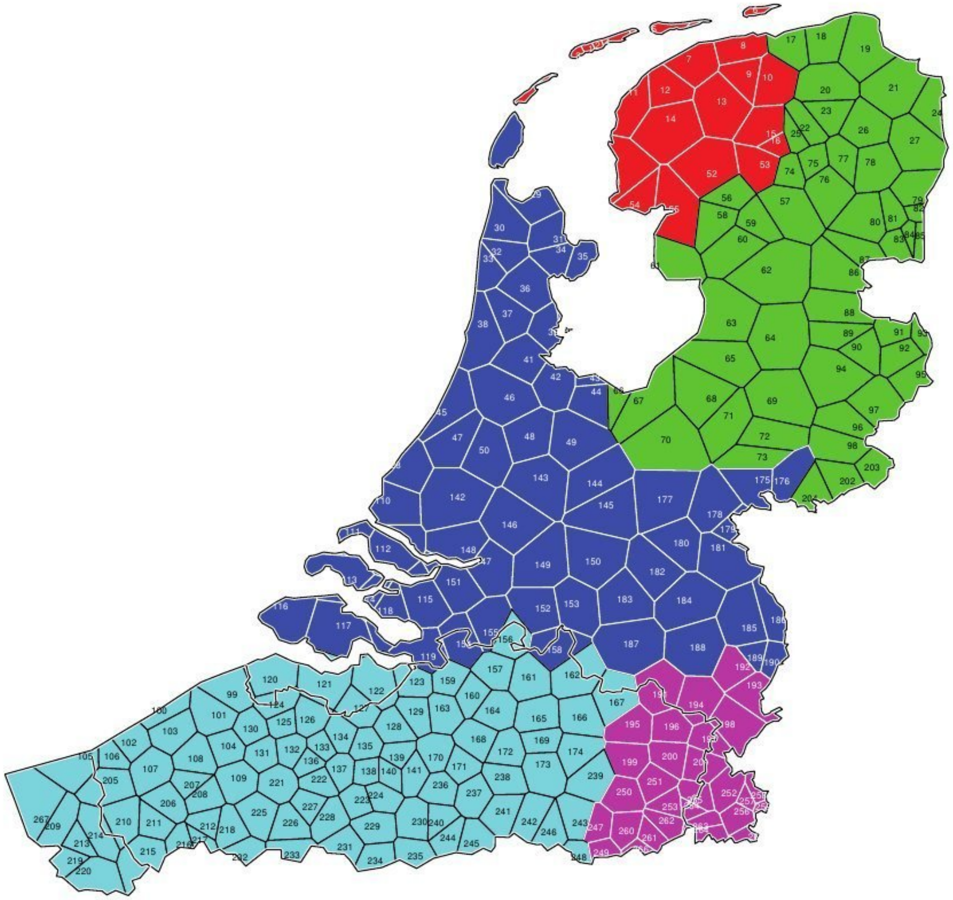
All values associated with a location are combined with the weights model to estimate their most likely region

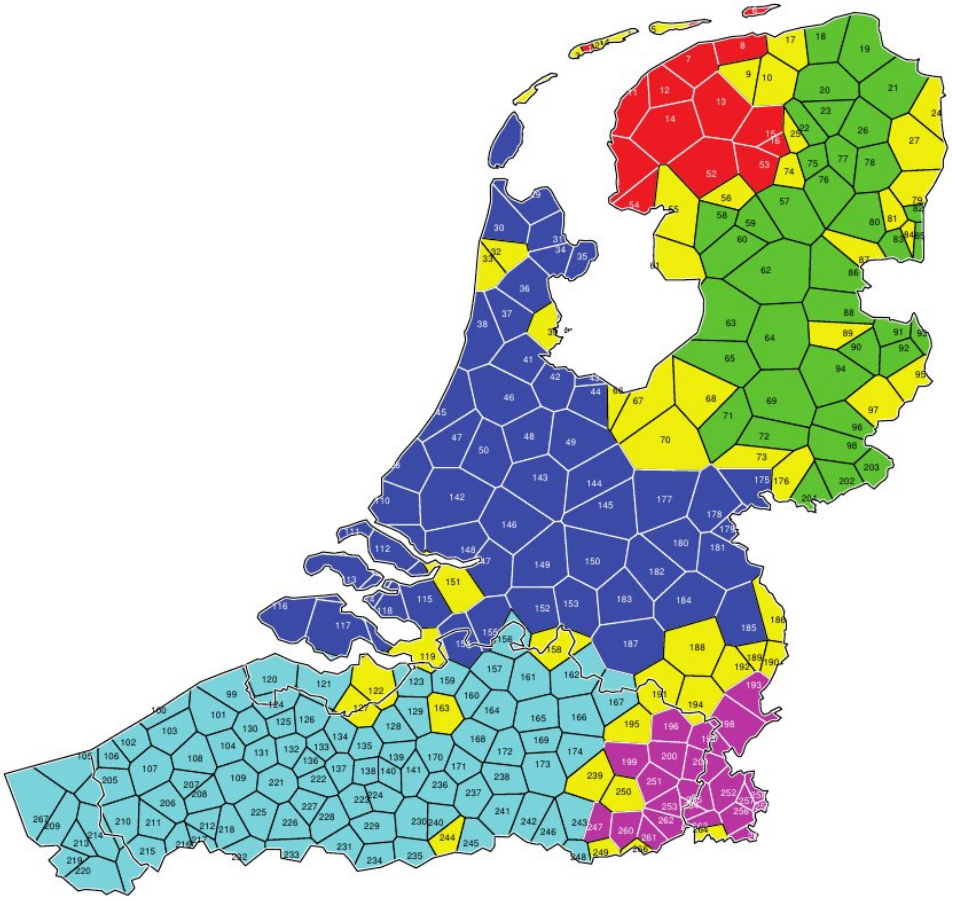
Results

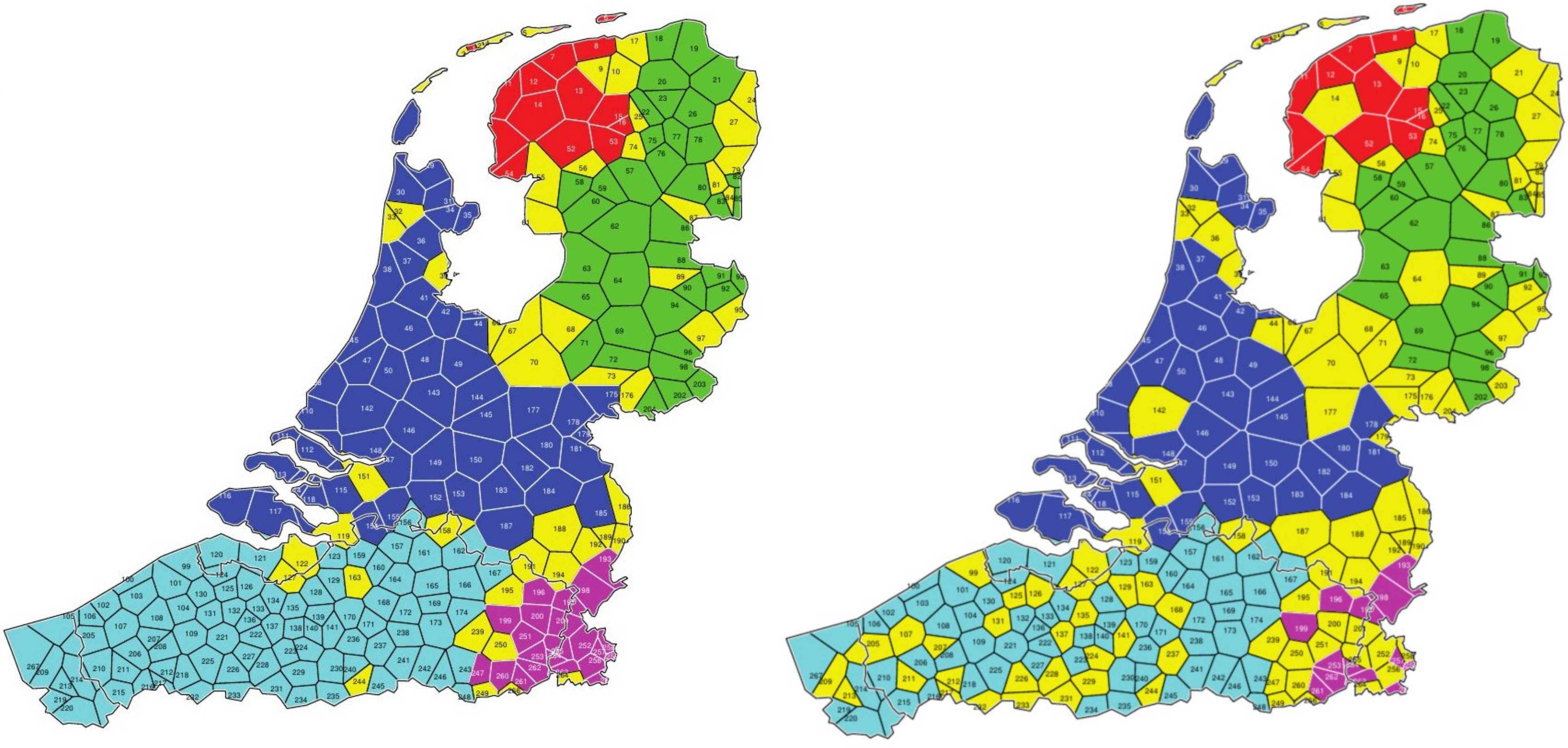
The Naive Bayes assigned the correct location to 217 of the 267 locations (81%)

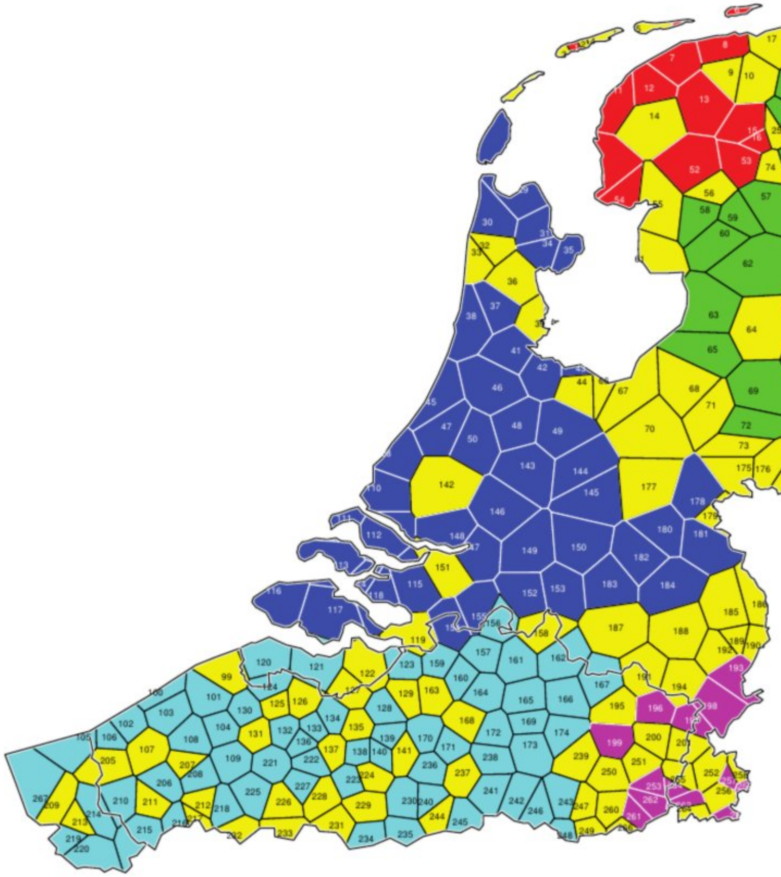
The 50 locations that were misclassified were mostly located along dialect boundaries

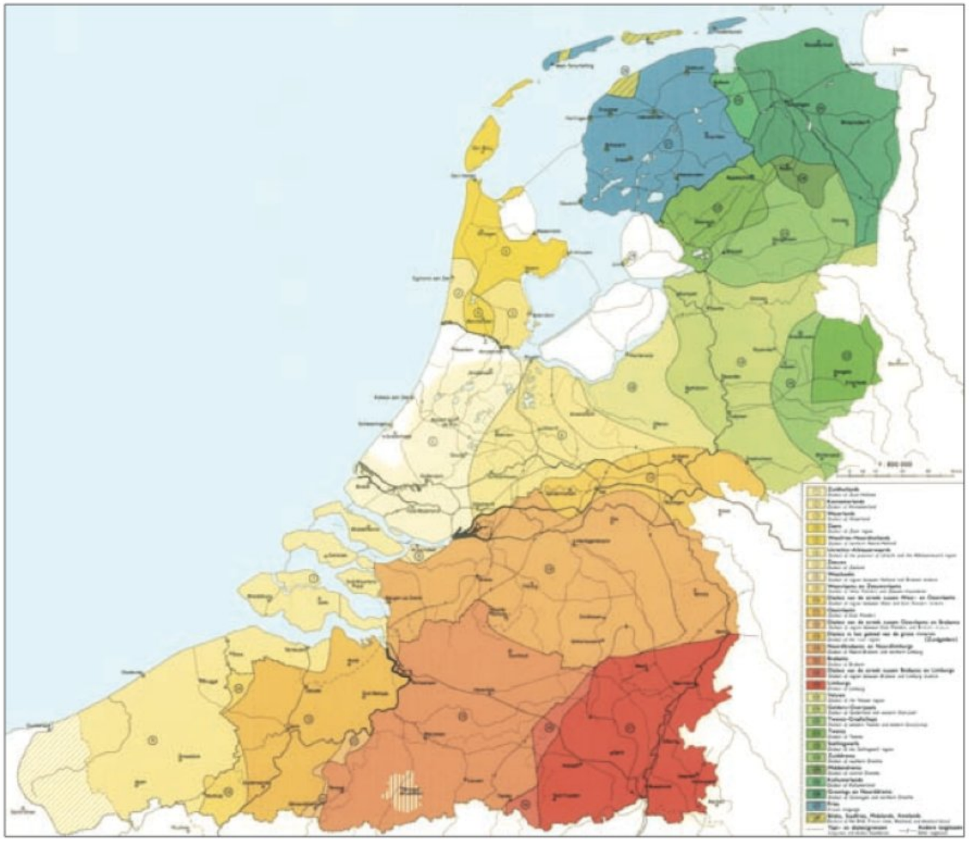
These misclassified locations are good candidates for transition zones: intermediate regions which share characteristics of neighboring regions


7
8
17
18
19
12
13
9
10
20
21
14
23
24
15
16
25
22
26
27
52
53
74
75
77
78
76
54
55
56
57
79
82
58
59
80
81
83
84
85
29
30
31
34
35
32
33
60
62
87
86
36
61
37
38
39
63
64
88
89
91
93
90
92
41
42
65
43
44
66
94
95
46
67
68
69
45
71
97
47
48
49
50
70
72
96
98
73
203
143
144
175
176
202
204
142
177
110
145
178
146
179
180
111
112
148
181
147
149
150
182
113
151
115
152
153
183
184
116
118
117
185
186
155
156
187
188
154
158
119
189
192
190
120
121
157
161
162
123
159
193
99
122
160
124
127
194
100
101
125
126
129
163
164
165
166
167
195
196
197
198
103
130
128
102
134
168
104
133
131
132
135
139
170
169
172
174
105
106
108
136
199
200
201
107
137
138
140
141
171
173
221
222
109
239
205
207
208
236
238
206
223
224
237
250
251
252
227
225
226
228
241
253
267
210
211
212
218
229
230
240
242
243
262
257
209
217
244
245
246
247
260
261
213
214
216
231
248
249
215
232
233
234
235
219
220


7
8
17
18
19
9
10
12
13
20
21
14
23
24
15
16
25
22
26
27
52
53
74
75
77
78
76
54
55
56
57
79
82
58
59
80
81
83
84
85
29
30
31
34
35
32
33
60
62
87
86
36
61
37
38
39
63
64
88
89
91
93
90
92
41
42
65
43
44
66
94
95
46
67
68
69
45
71
97
47
48
49
50
70
72
96
98
73
203
143
144
175
176
202
204
142
177
110
145
178
146
179
180
111
112
148
181
147
149
150
182
113
151
115
152
153
183
184
116
118
117
185
186
155
156
187
188
154
158
119
189
192
190
120
121
157
161
162
123
159
193
99
122
160
124
127
194
100
101
125
126
129
163
164
165
166
167
191
195
196
197
198
103
130
128
102
134
168
104
133
131
132
135
139
170
169
172
174
105
106
108
136
199
200
201
107
137
138
140
141
171
173
221
222
109
239
205
207
208
236
238
206
223
224
237
250
251
252
227
225
226
228
241
253
267
210
211
212
218
229
230
240
242
243
262
257
209
217
244
245
246
247
260
261
213
214
216
231
248
249
215
232
233
234
235
219
220







Concluding remarks

We have presented results of experiments in finding dialect transition areas based on data from the Syntactic Atlas for Dutch Dialects

The interpretation of the data set proved to be nontrivial because of the lack of explicit negative information and a large number of missing values

An experiment with predicting regions based on pronunciation data showed that misclassification occurred mostly along dialect borders

The misclassified locations are good candidates for so-called transition zones

Future work

Translate results to something usable by linguists: which linguistic variables play a role in separating regions or defining transition zones?

Repeat experiment with *all* SAND variables: does it produce similar results

Find more general visualization software, ideally linked to existing map services

THE END