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 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?

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

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 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
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