Morphology and dialectology in the Linguistic Survey of Scotland

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Morphology and dialectology in the Linguistic Survey of Scotland

A quantitative approach

Pavel Iosad
pavel.iosad@ed.ac.uk
Will Lamb
w.lamb@ed.ac.uk
Oilthigh Dhùn Èideann
Rannsachadh na Gàidhlig
Sabhal Mòr Ostaig
24th June 2016

Outline

- Motivation
  - State of the art
  - What is dialectometry?
  - Why dialectometry?
- LSS(G) data
- Three different analyses
  - Spatial analysis
  - Correlation analysis of features
  - Correlation analysis of varieties
- Conclusions

1 Background and motivation

1.1 State of the art

Current status

- Individual dialect descriptions
The division of Gaelic dialects

- Many scholars have made comments on dialectal divisions in Gaelic
- The approach is either purely historical (e.g., Jackson) or impressionistic
- No solid data:
  - SGDS exists for qualitative analysis, but not much work has been done with it
  - No quantitative data

1.2 Dialectometric approach

What is dialectometry?

‘Dialectometry studies dialects using exact methods, especially computational and statistical approaches’ (Wieling & Nerbonne 2015)

- Focus on objective, quantitative methods
- Focus on aggregate measures not individual features
- ‘Individual features are inevitably noisy’
- Covers both spatial variation and variation within a location
Common methods

- String distance (e.g. Levenshtein distance)
- Clustering methods (e.g. Ward clustering)
- Multidimensional analysis
- Correlation analysis
- Regression (including spatially adjusted methods)

Common applications

- Pronunciation distance
- Cluster analysis: alternative to traditional isoglosses
- Multidimensional analysis: identifying dialect areas from the data
- Mostly based on phonetic material!
- Wieling & Nerbonne (2015): not much has been done on morphosyntax, though increasing interest in recent years

Previous applications to Celtic

- Recent reevaluation for Irish by Ó Muircheartaigh (2014)
- Some work on Breton, see Brun-Trigaud, Sollec & Le Dû (2016) with references

2 Data

2.1 LSS morphology data

Linguistic Survey: background

- Main collection period: 1951–63
  - Coverage very close to 18th century ‘Highland Line’
  - Impressive given Jackson’s famously strict criteria
- Questionnaire sections
  - Phonology: 893 headwords
    * Published as Ó Dochartaigh (1994–1997)
  - Morphophonology and syntax
    * 13.5 pages, unpublished
Example materials

2.2 Our study

Coding

• Coded by hand from original field materials at the School of Scottish Studies Archives
  -- 1 for presence of feature
  -- 0 for absence of feature
  -- Blank for no return

• Features coded using target phrase, asterisk marks feature of interest
  • E.g. na casan beag*: presence of suffix in feminine plural adjectives
    -- 1 for na casan beaga
    -- 0 for na casan beag or any other form

• Ongoing: mapping demographic data reporting in the LSS to census return to evaluate potential effects of language shift/obsolescence

Analysis

• All analysis conducted with R (R Core Team 2016)
• Methods
  -- Generalized additive models with package mgcv (Wood 2006)
  -- Cluster analysis with package cluster (Maechler et al. 2015)
  -- Correlation analysis with R core function cor and corrplot package (Wei & Simko 2016)
3 Results

3.1 Spatial variation

Method

- Logistic regression: probability of feature being present depending on latitude and longitude
- Non-linear regression: generalized additive models (Wood 2006)

Current more a visualization method than a predictive analysis
- But can be combined with explanatory variables to adjust for them: current plan to do this with demographic data

Local pattern

dà [mhnaoi]
The cline

+ fhir bh*ig

- The smoothing allows us to see the 'big picture'
- There is a southeast-northwest cline
- Could be related to language decline?

Next steps: include demographic data as explanatory variable to adjust for it.
The 'Uist-Barra' effect

leis a' chois bh*ig

don [bhoin]
3.2 Correlation and clustering: dialects

Correlation analysis

- We can represent each dialect as a sequence of values (a vector)

<table>
<thead>
<tr>
<th>ID</th>
<th>Port of Ness</th>
<th>Upper Shader</th>
<th>Bragar</th>
<th>Carloway</th>
<th>Brenish</th>
<th>North Tolsta</th>
<th>Lower Pabail</th>
<th>Leurbost</th>
<th>Gravir</th>
<th>Scarp</th>
<th>Archasaig</th>
<th>Grosebay</th>
<th>Leverburgh</th>
<th>St Kilda 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- Port of Ness = \(<1, 1, 1 \ldots>\)
- We can calculate the correlation matrix for a set of vectors
- The higher the correlation, the more similar the dialects are to each other
- A correlation of 1 means their behaviour is identical, a correlation of −1 means they are exact opposites

Cluster analysis

- Once we have a correlation matrix, we can rank the dialects in terms of how close they are to each other
- Based on this, we are able to conduct clustering
- Various methods: agglomerative Ward clustering is common
- We set the number of cuts to make in the tree
- Here: three clusters
Results

- Confirms some qualitative observations:
  - Cluster 3 (green): concentrated in Uist/Barra/Harris
– Cluster 2 (blue): periphery (correlation with strength of Gaelic?)
  • More fine-grained analysis also possible

3.3 Correlation analysis: features

Correlation of features

• We can use the same technique to evaluate how similar the features are across dialects
• This can tell us about patterns of changes (and obsolescence)
• Adger (2016) suggests that simultaneous changes in apparently unrelated aspects of grammar may reveal the underlying unity of the grammatical mechanisms involved
Genitive articles

- A set of correlated features is the use of *na* in the genitive
  - *[na]* súla glaise
  - *[na]* cathrach bige
  - *[na]* coise bige

- Methodological sanity check
  - Different feminine lexical items lose the genitive form of the article together
  - Candidate for least surprising finding of the year, but this shows our data and methods produce at least some plausible results

Loss of lenition

- One very clear cluster is formed by ‘core’ lenition contexts:
  - *a’ chas bhéag*
  - *a’ chas*
  - *an fhír*

- Lenition in these three contexts is lost simultaneously (in diatopic terms)
- But: no correlation with loss of lenition in some other contexts (e.g. *(a) fhír bhíg)*
- No single grammatical mechanism for *all* lenition
- The simultaneity in these three contexts could show that they do reflect a single underlying mechanism

See Iosad (2014) for similar reasoning on Breton spirantization

3.4 Conclusions and prospects

Conclusions

- A quantitative approach to Gaelic dialectology is possible and worthwhile
  - Produces plausible results
  - Allows us to ask new questions

- Potential for insights into diatopic variation beyond ‘centre and periphery’, with adjustment for other factors
- Potential for analytic insights into linguistic structure

Prospects

- Limitation of coding: currently all 0 cells are equal (count for similarity calculations) even if the forms are not identical

This would need more detailed coding, but for many of our variables it doesn’t really matter
- Add explanatory variables
- Combine with phonetic data (SGDS): stay tuned!
- Use insights gained to calibrate traditional/anecdotal knowledge of morphosyntactic variation: important for corpus planning (Bell et al. 2014)
References

Adger, David. 2016. Structure, use, and syntactic ecology in language obsolescence. MS., Queen Mary University of London.
Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert & Kurt Hornik. 2015. cluster: Cluster Analysis Basics and Extensions. Version 2.0.3.
Appendix: Jackson on fieldwork

- On change
  - ‘Remarkable that AF’s case system is so much more decayed [...] though he is the same age as JM. AM has kept it well’

- On omissions
  - ‘Really this sort of thing would try the patience of a saint. Particularly since Barrai is especially interesting in preserving the forms of the adjective rather well!’

- On informants
  - ‘An ideal informant, a first-rate mind with natural flair for analysis. Hardly literate in Gaelic. Does not now use Gaelic much actively.’
  - ‘Struck me as a crude and uneducated old man but this questionnaire suggests rather that he knows written Gaelic.’
  - ‘[an] ideal informant: totally unsophisticated’