Mapping syntactic variation in varieties of English world-wide

A comparative sociolinguistic approach

Melanie Röthlisberger
melanie.roethlisberger@es.uzh.ch
Introduction
Theoretical background

Two research paradigms:

- **Probabilistic grammar** (e.g. Bresnan 2007)
  - grammatical knowledge is partially probabilistic
  - multiple probabilistic constraints influence the outcome of grammatical variation
  - grammatical knowledge is experience- and usage-based

- **World Englishes** (e.g. Schneider 2007)
  - structural characteristics and sociohistorical background of varieties of English
Background: World Englishes

“indigenization of language structure mostly occurs at the interface between grammar and lexis, affecting the syntactic behavior of certain lexical elements. Individual words, typically high frequency items, adopt characteristic but marked usage and complementation patterns.”

(Schneider 2007: 46; emphasis mine)

(1) no worries
(2) this hair-style is called as ‘duck tail’
Background: Probabilistic grammar

Indigenization not only occurs at the lexis-syntax interface but also with regard to the probabilistic constraints that influence linguistic variation (e.g. end-weight)

Mary gave [John] [the apple]  Mary gave [the apple] to [John]
Background: Probabilistic grammar

probabilistic indigenization =

“the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A [...] can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A.”

(Szmrecsanyi et al. 2016: 133)
Aim

• How different/similar are varieties with regard to the stochastic constraints that shape linguistic variation?

• To what extent do traditional classifications in dialect typology correspond to varieties’ classification based on the underlying probabilistic grammar?
Methodology & data
Methodological steps

1. Identify a linguistic variable
2. Select appropriate data/corpus
3. Extract variable tokens from the data
4. Annotate each token for probabilistic constraints (e.g. length of constituents)
5. Fit statistical model per variety to gauge the influence of factors
6. Use comparative sociolinguistic approach to quantify dissimilarity between varieties
7. Visualize distances between varieties (MDS, cluster)
1. Linguistic variable

The English dative alternation:

(3) ditransitive dative

Mary gives [John]_{recipient} [an apple]_{theme}

(4) prepositional dative

Mary gives [an apple]_{theme} to [John]_{recipient}
2. Corpus data

Canadian English
British English
Irish English
Jamaican English
Indian English
Singapore English
New Zealand English
Hong Kong English
Philippine E.
2. Corpus data

- International Corpus of English (ICE) - series
  - 60% spoken (transcriptions), 40% written texts = 1m words per subcorpus
    - 500 texts, 2,000 words per text
    - 12 different registers, same corpus structure

- Corpus of Global web-based English (GloWbE)
  - general websites and blogs
  - data sampled in 2012-13 – https://corpus.byu.edu/glowbe/
3. Extract and 4. Annotate

(e.g. Bresnan et al. 2007)

- retrieval of dative variants using verb list and perl script
- restrict to choice context (incl. pronouns)
- code for numerous (language-internal) factors: length (weight ratio), complexity, pronominality, givenness, definiteness, person, animacy, concreteness of theme, verb sense
- code for language-external factors: Mode (spoken vs written)

\[ N = 13,171 \]
5. Statistical modeling

• mixed-effects logistic regression analysis to gauge the influence of all probabilistic constraints simultaneously while taking idiosyncrasies of the data into account (e.g. lexical effects, corpus structure, speakers)

• conditional random forests to identify the relative importance of constraints per variety (ranking)

• R: lme4, party packages

⇒ fit one model per variety and compare
6. Comparative sociolinguistic approach

“Three lines of evidence” to compare dialects / varieties in the probabilistic domain (Poplack and Tagliamonte 2001)

1. statistical significance ⇒ mixed-effect
2. relative strength ⇒ mixed-effect
3. Constraint ranking ⇒ random forest
6. Comparative sociolinguistic approach

1. Fit a mixed-effects model /conditional random forest per variety using the same model formula

2. Calculate / Create a distance matrix:
   - **statistical significance**: compare number of shared significant and non-significant constraints
   - **relative strength**: use Euclidean distance metric to calculate distance between coefficient estimates from models
   - **constraint ranking**: calculate Spearman’s rank correlation coefficient between the constraint ranks as a distance measure

3. Visualize: reduce number of dimensions with Multidimensional scaling (Kruskal and Wish 1978) and/or cluster based on similarities/differences
6. Comparative sociolinguistic approach

**1st line:** Statistical significance

**2nd line:** Relative strength

Random effects:
- Groups: FileID, ThemeHeadFilter, Verb, RecHeadFilter
- Variance: 5.371e-13, 2.735e+00, 2.411e+00, 0.000e+00
- Std.Dev.: 7.329e-07, 1.654e+00, 1.553e+00, 0.000e+00

Number of obs: 1318, groups: FileID, 571; ThemeHeadFilter, 44; Verb, 29

Fixed effects:
- Estimate: -4.33676, 2.78005, 1.31865, 1.08568, -0.04476, 0.87375, 0.81576, 0.56493, 0.03603, 1.29943, 0.95305
- Std. Error: 0.66134, 0.45372, 0.37445, 0.31880, 0.33245, 0.25698, 0.37352, 0.29037, 0.25528, 0.42333, 0.25923
- z value: -6.558, 6.127, 3.522, 3.400, -0.073, 3.400, 2.184, 1.946, 0.143, 3.070, 3.676
- P value: 5.47e-11, 8.94e-10, 0.000429, 0.000660, 0.943573, 0.000674, 0.028964, 0.051705, 0.887750, 0.002144, 0.000237

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
6. Comparative sociolinguistic approach

- LengthOfObject
- SemanticsOfObject
- SemanticsOfSubject
- GivennessDefPron
- LengthOfSubject
- Variety
- VPcomplexity

3rd line: constraint ranking

1. 
2. 
3. 
4. 
5. 
6. 
7.
### Example: Relative strength

<table>
<thead>
<tr>
<th>Factor</th>
<th>GB</th>
<th>CAN</th>
<th>IRE</th>
<th>NZ</th>
<th>JA</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-4.337</td>
<td>-5.027</td>
<td>-4.662</td>
<td>-3.469</td>
<td>-3.519</td>
<td></td>
</tr>
<tr>
<td>Weight-ratio</td>
<td>2.780</td>
<td>3.376</td>
<td>2.203</td>
<td>2.430</td>
<td>4.440</td>
<td>...</td>
</tr>
<tr>
<td>RecPron</td>
<td>1.319</td>
<td>2.651</td>
<td>2.717</td>
<td>2.229</td>
<td>2.343</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
## Example: Relative strength

<table>
<thead>
<tr>
<th></th>
<th>GB</th>
<th>CAN</th>
<th>IRE</th>
<th>NZ</th>
<th>JA</th>
<th>SIN</th>
<th>IND</th>
<th>HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAN</td>
<td>3.218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRE</td>
<td>2.452</td>
<td>2.314</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NZ</td>
<td>2.314</td>
<td>2.090</td>
<td>1.674</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JA</td>
<td>3.142</td>
<td>3.920</td>
<td>3.939</td>
<td>3.420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIN</td>
<td>2.224</td>
<td>2.173</td>
<td>1.461</td>
<td>0.682</td>
<td>3.798</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IND</td>
<td>2.245</td>
<td>2.208</td>
<td>1.343</td>
<td>1.830</td>
<td>3.581</td>
<td>1.670</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>1.508</td>
<td>2.036</td>
<td>1.789</td>
<td>1.695</td>
<td>3.118</td>
<td>1.648</td>
<td>1.682</td>
<td></td>
</tr>
<tr>
<td>PHI</td>
<td>3.552</td>
<td>1.594</td>
<td>2.607</td>
<td>2.300</td>
<td>4.402</td>
<td>2.425</td>
<td>2.732</td>
<td>2.576</td>
</tr>
</tbody>
</table>
Example: Relative strength

![Graph showing relative strength with variances](image)
Results

Quantifying the syntactic distance between varieties
Comparing significance (1st line)
Comparing relative strength (2nd line)
Comparing constraint hierarchy (3rd line)
## Clustering variety types: Relative strength

<table>
<thead>
<tr>
<th></th>
<th>GB</th>
<th>CAN</th>
<th>IRE</th>
<th>NZ</th>
<th>JA</th>
<th>SIN</th>
<th>IND</th>
<th>HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAN</td>
<td>3.218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRE</td>
<td>2.452</td>
<td>2.314</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NZ</td>
<td>2.314</td>
<td>2.090</td>
<td>1.674</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JA</td>
<td>3.142</td>
<td>3.920</td>
<td>3.939</td>
<td>3.420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIN</td>
<td>2.224</td>
<td>2.173</td>
<td>1.461</td>
<td>0.682</td>
<td>3.798</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IND</td>
<td>2.245</td>
<td>2.208</td>
<td>1.343</td>
<td>1.830</td>
<td>3.581</td>
<td>1.670</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HK</td>
<td>1.508</td>
<td>2.036</td>
<td>1.789</td>
<td>1.695</td>
<td>3.118</td>
<td>1.648</td>
<td>1.682</td>
<td></td>
</tr>
<tr>
<td>PHI</td>
<td>3.552</td>
<td>1.594</td>
<td>2.607</td>
<td>2.300</td>
<td>4.402</td>
<td>2.425</td>
<td>2.732</td>
<td>2.576</td>
</tr>
</tbody>
</table>
Clustering variety types

Hierarchical clustering of by-variety models (using ward)

varieties
hclust (*, "ward.D2")
Discussion & Conclusion
Discussion

- How different/similar are varieties with regard to the stochastic constraints that shape linguistic variation?
  ⇒ Visualizations of the probabilistic distances have highlighted shared probabilistic grammar

- To what extent do traditional classifications in dialect typology correspond to varieties’ classification based on the underlying probabilistic grammar?
  ⇒ No L1 vs L2 distinction but American vs. British-based variety
Conclusion

Reconsider the existing dichotomies of variety types (ENL – ESL – EFL), the Dynamic Model, etc. by

• using a bottom-up approach
• going beyond pure frequencies
• establishing linguistic similarities between varieties based on all levels of grammar (more than just one variable)
Thank you!

melanie.roethlisberger@es.uzh.ch

www.melanie-roethlisberger.ch


