Measuring the variational homogeneity of English as a world language: Probabilistic indigenization effects in four syntactic alternations

Iván Tamaredo, University of Santiago de Compostela
Benedikt Szmrecsanyi, KU Leuven
Jason Grafmiller, University of Birmingham
Benedikt Heller, Justus Liebig University of Giessen
Melanie Röthlisberger, University of Zurich
Theoretical background

Two research paradigms:

- Probabilistic grammar (e.g., Bresnan 2007)
- World Englishes (e.g., Schneider 2007)

1. grammatical knowledge is partially probabilistic
2. multiple probabilistic constraints influence the outcome of grammatical variation
3. grammatical knowledge is experience- and usage-based

How similar or dissimilar is the probabilistic knowledge of English grammar on the part of speakers with different regional and cultural backgrounds?
Theoretical background

• Varieties share a common probabilistic grammar in that some language-internal constraints are largely shared (e.g., Bernaisch et al. 2014; Szmrecsanyi et al. 2016).

• BUT some constraints impact a speaker’s choice differently in one variety compared to another.
  - Example: LENGTH OF CONSTITUENTS has a weaker impact in variety A than in variety B on the choice between Mary gives John the apple vs. Mary gives the apple to John

• Indigenization: “the emergence of locally characteristic linguistic patterns” (Schneider 2007: 6).
  - Lexical items in novel syntactic constructions: e.g., visit with in Philippine English instead of visit.
Theoretical background

• Probabilistic indigenization:
  “the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties” (Szmrecsanyi et al. 2016: 133)

• Hypothesis: probabilistic indigenization effects arise as a function of the lexical specificity of alternations, with those that are strongly connected with specific lexical items being the most likely ones to exhibit cross-varietal indigenization effects.
Aims

• Measure the degree of alternation-internal homogeneity or heterogeneity across three varieties: British English, Indian English, and Singapore English.

• How speakers select a specific variant when they have a choice between “alternate ways of saying ‘the same’ thing” (Labov 1972: 188).

• Four syntactic alternations: dative \((N=3,012;\) see, e.g., Bresnan & Hay 2008), genitive \((N=3,108;\) see, e.g., Rosenbach 2014), particle placement \((N=2,480,\) see, e.g., Gries 2003), and subject pronoun omission \((N=2,456;\) see, e.g., Torres Cacoullos & Travis 2014)
Aims

(1) a. The ditransitive dative variant

    That will give [the panel]$_{\text{recipient}}$ [a chance]$_{\text{theme}}$ to expand on what they’ve been saying. (ICE-GB:S1B-036)

b. The prepositional dative variant

    [...] and that gives [a chance]$_{\text{theme}}$ [to Bhupathy]$_{\text{recipient}}$ to equalise the points at thirty all. (ICE-IND:S2A:019)
Aims

(2)  

a. The s-genitive

\[ \text{Singapore}_\text{possessor's} \ [\text{small size}_\text{possessum} \text{ meant it could be quick to respond to changes in economic conditions} \ (\text{ICE-SIN:W2C-011}) \]

b. The of-genitive

\[ \text{the} \ [\text{size}_\text{possessum} \text{ of} \ [\text{the eyes}_\text{possessor} \text{ is to help them at night}. \ (\text{ICE-GB:W2B-021}) \]
Aims

(3)  a. Verb-object-particle order (or split order)

*you can just [cut]_verb [the tops]_direct object [off]_particle and leave them.*
(ICE-GB:S1A-007)

b. Verb-particle-object order (or joined order)

*[Cut]_verb [off]_particle [the flowers]_direct object as they fade.* (ICE-CAN:W2B-023)
Aims

(4) a. Overt subject pronoun

*The vision$_i$ was not very clear. It$_i$ was murky or rather uh foggy or misty.* (ICE-IND:S1B-006)

b. Null subject pronoun

*Oh, be4 I forget, “Chitra$_i$” sends you her love. Ø$_i$ Has been asking about you since you left.* (ICE-SIN:W1B-003)
Aims

The greater the degree of probabilistic indigenization (i.e. the smaller varieties’ similarity)

The greater the impact of lexical specific constituents.
Data & methodology

• Relevant observations of the (a) and (b) variants of the four alternations retrieved from the British, Indian, and Singaporean components of the International Corpus of English (ICE).

• 5 most important predictors selected on the basis of conditional random forests fitted to the dataset of all three varieties.
## Data & methodology

<table>
<thead>
<tr>
<th>PREDICTORS DATIVE</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight ratio</td>
<td>Recipient length in letters divided by theme length in letters (log value)</td>
</tr>
<tr>
<td>Recipient pronominality</td>
<td>Pronoun vs nominal</td>
</tr>
<tr>
<td>Recipient person</td>
<td>Local vs non-local</td>
</tr>
<tr>
<td>Theme complexity</td>
<td>Simple vs complex</td>
</tr>
<tr>
<td>Recipient head frequency</td>
<td>Global text frequency of recipient head (lemma)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PREDICTORS GENITIVE</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possessor animacy</td>
<td>Human &amp; animal vs collective vs inanimate vs locative vs temporal</td>
</tr>
<tr>
<td>Possessor length</td>
<td>Length of possessor in letters</td>
</tr>
<tr>
<td>Possessum length</td>
<td>Length of possessum in letters</td>
</tr>
<tr>
<td>Possessor thematicity</td>
<td>Number of uses of the possessor head noun in a text divided by the total number of words in the text</td>
</tr>
<tr>
<td>Possessor final sibilancy</td>
<td>Yes vs no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PREDICTORS PART. PL.</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct object length</td>
<td>Length of direct object in words</td>
</tr>
<tr>
<td>Semantics</td>
<td>Compositional vs non-compositional</td>
</tr>
<tr>
<td>Directional PP</td>
<td>Yes vs no</td>
</tr>
<tr>
<td>Verb surprisal</td>
<td>Predictability of the verb given the particle</td>
</tr>
<tr>
<td>Preposition surprisal</td>
<td>Predictability of the particle given the verb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PREDICTORS SUBJ. OM.</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text type</td>
<td>Spoken informal vs spoken formal vs written informal vs written formal</td>
</tr>
<tr>
<td>Coordination</td>
<td>Coordination vs no coordination</td>
</tr>
<tr>
<td>Clause type</td>
<td>Main vs embedded</td>
</tr>
<tr>
<td>Clause position</td>
<td>Initial vs non-initial</td>
</tr>
<tr>
<td>Pronoun-verb cooccurrence frequency</td>
<td>How many times the pronoun and the verb cooccur</td>
</tr>
</tbody>
</table>
Data & methodology

• Per variety binary mixed-effects logistic regression and conditional random forest analyses:
  ➢ mixed-effects models included random intercepts for lexical items.

• Comparative sociolinguistics (Poplack & Tagliamonte 2001, *inter alii*):
  ➢ compares and contrasts patterns of variability of linguistic features across different dialects or varieties using quantitative methods.

• Three lines of evidence:
  1. shared significant/non-significant predictors
  2. relative strength of predictors
  3. importance or rank of predictors
Data & methodology

• Three steps:
  1. Fit a mixed-effects model /conditional random forest per variety using the same model formula per alternation.
  2. Calculate a distance matrix:
     a. **statistical significance**: number of shared significant and non-significant predictors (mixed-effects models)
     b. **relative strength**: distance between coefficient estimates from models (mixed-effects models)
     c. **constraint ranking**: Spearman’s rank correlation coefficient between the constraint ranks as a distance measure (conditional random forests)
  3. Calculate the average similarity as a measure of overall stability.

• Three core grammar coefficients (0-1): the higher the value, the more homogeneous the alternation.
Data & methodology

• To gauge lexical effects: visualize the random slopes of the mixed-effects models.

• Lemmas of individual lexical items as random effects:
  - verbs, recipients, and themes in the dative alternation
  - possessors and possessums in the genitive alternation
  - verbs, particles, and verb-particle combinations in the particle placement alternation
  - verbs in the case of subject omission

• Variance accounted for by lexical effects in the random structure of the mixed-effects model: \( \text{r.squaredGLMM()} \) in MuMIn package (Barton 2015).

• R\(^2\) provides indication of model fit
  - Marginal R\(^2\) = variance accounted for in model with fixef
  - Conditional R\(^2\) = variance accounted for in model with ranef + fixef
  - variance accounted for by lexical effects (ranef) only: cR\(^2\)−mR\(^2\)
Results: core grammar coefficients

<table>
<thead>
<tr>
<th></th>
<th>Datives</th>
<th>Genitives</th>
<th>Particles</th>
<th>Subject o.</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance</td>
<td>0.867</td>
<td>1.000</td>
<td>0.600</td>
<td>0.619</td>
<td>0.772</td>
</tr>
<tr>
<td>Effect strength</td>
<td>0.593</td>
<td>0.531</td>
<td>0.073</td>
<td>0.787</td>
<td>0.496</td>
</tr>
<tr>
<td>Constraint ranking</td>
<td>0.733</td>
<td>0.833</td>
<td>0.733</td>
<td>0.800</td>
<td>0.775</td>
</tr>
<tr>
<td>Mean</td>
<td>0.731</td>
<td>0.788</td>
<td>0.469</td>
<td>0.721</td>
<td>0.677</td>
</tr>
</tbody>
</table>

- Global mean: overall, varieties seem to share a core probabilistic grammar
- Mean per line: constraint ranking > significance > effect strength
- Mean per alternation: genitives > datives > subject omission > particle placement
Results: lexical specificity

• Dative (recipient not shown as SD=0)
Results: lexical specificity

- Genitive
Results: lexical specificity

• Particle placement
Results: lexical specificity

• Subject omission
Results: lexical specificity

• Lexical specificity (from most to least):
  ➢ genitives > datives/particles > subjects
  ➢ no cross-varietal patterns

• Homogeneity:
  genitives > datives > subjects > particles

<table>
<thead>
<tr>
<th>Alternation</th>
<th>Marginal R²</th>
<th>Conditional R²</th>
<th>% of variance accounted for by random structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dative alternation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BrE</td>
<td>0.433</td>
<td>0.805</td>
<td>0.372</td>
</tr>
<tr>
<td>IndE</td>
<td>0.087</td>
<td>0.173</td>
<td>0.086</td>
</tr>
<tr>
<td>SinE</td>
<td>0.209</td>
<td>0.549</td>
<td>0.34</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td></td>
<td>0.266</td>
</tr>
<tr>
<td>Genitive alternation:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BrE</td>
<td>0.293</td>
<td>0.762</td>
<td>0.469</td>
</tr>
<tr>
<td>IndE</td>
<td>0.409</td>
<td>0.666</td>
<td>0.257</td>
</tr>
<tr>
<td>SinE</td>
<td>0.431</td>
<td>0.717</td>
<td>0.286</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td></td>
<td>0.337</td>
</tr>
<tr>
<td>Particle placement:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BrE</td>
<td>0.324</td>
<td>0.492</td>
<td>0.168</td>
</tr>
<tr>
<td>IndE</td>
<td>0.215</td>
<td>0.551</td>
<td>0.336</td>
</tr>
<tr>
<td>SinE</td>
<td>0.324</td>
<td>0.617</td>
<td>0.293</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td></td>
<td>0.266</td>
</tr>
<tr>
<td>Subject omission:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BrE</td>
<td>0.749</td>
<td>0.782</td>
<td>0.033</td>
</tr>
<tr>
<td>IndE</td>
<td>0.596</td>
<td>0.709</td>
<td>0.113</td>
</tr>
<tr>
<td>SinE</td>
<td>0.568</td>
<td>0.618</td>
<td>0.05</td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
<td></td>
<td>0.065</td>
</tr>
</tbody>
</table>
Summary

- **Genitive alternation**
  - *most stable*

- **Dative alternation**

- **Subject omission**

- **Particle placement**
  - *least stable*

- **Subject omission**

- **Particle placement / Dative alternation**

- **Genitive alternation**
  - *least lexical specific*

- **Most lexical specific**
Conclusions

• Probabilistic indigenization can be observed to different degrees in the three varieties and across four alternations

• Various explanations have been offered (see e.g. Röthlisberger et al. 2017)

• The degree of alternation-internal homogeneity is not directly linked to an alternation’s lexical specificity but seems to be reversed
  ➢ genitive alternation exhibits the most stability in probabilistic constraints but is also the most lexical specific by variety
  ➢ it seems not to be the case that the degree of probabilistic indigenization can be linked to an alternation’s lexical specificity
Conclusions

The **greater** the degree of probabilistic indigenization (i.e. the smaller varieties’ similarity)

The **greater** the impact of lexical specific constituents.
Next steps

• $R^2$ might not be a good heuristic to assess the lexical specificity of an alternation $\rightarrow$ collostructional analysis

• use other measures to compare/contrast varieties (e.g. AIC see Grafmiller & Szmrecsanyi under revision)
KIITOS

Slides can be downloaded from: www.melanie-roethlisberger.ch/research/publications
References

Primary sources:


Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). https://corpus.byu.edu/glowbe

International Corpus of English – the British Component. 1998. Project coordinated by Bas Aarts at University College London, United Kingdom.


Secondary sources:


