Using character $n$-grams to classify native language in a non-native English corpus of transcribed speech

Charlotte Vaughn
Janet Pierrehumbert
Hannah Rohde

Northwestern University
Authorship attribution
(Mosteller and Wallace, 1964; Koppel, Schler, and Zigdon, 2005)

- Use various components of writing (e.g. syntactic, stylistic, discourse-level) to determine aspects of author’s identity
  - e.g. gender, emotional state, native language, actual identity
Native language classification
(Tsur and Rappoport, 2007)

- Examined English writing from the International Corpus of Learner English (ICLE)
  - Used subcorpora from 5 different native language backgrounds: Bulgarian, Czech, French, Russian, Spanish

- Divided each document into character $n$-grams
  - e.g. ‘bigrams’ = ‘_b’, ‘bi’, ‘ig’, ‘gr’, ‘ra’, ‘am’, ‘ms’, and ‘s_’

- Used multi-class support vector machine (SVM) to classify each document by native language of writer
Findings
(Tsur and Rappoport, 2007)

Obtained 65.6% accuracy in identifying native language of the author based on character bigrams alone.

- Compared with 20% random baseline accuracy, 46.78% accuracy for character unigrams, and 59.67% for character trigrams.
Speculated that “use of L2 words is strongly influenced by L1 sounds and sound patterns” (p. 16)
bigrams ≈ diphones

Language transfer evident on many levels
  – Effect of L1 on L2 pronunciation is widely attested
    (Flege, 1987, 1995; Mack, 2003)

But, what if your L1 background doesn’t just affect how you say words in your L2, but what words you use in the first place?
Drawbacks and open questions from Tsur and Rappoport (2007)

- How generalizable are these results to speech?
  - Writing is a more conscious, deliberate process than speech
  - If this really is a phonological process, we might expect stronger effects in speech

- Used corpus uncontrolled for topic content
  - Did use *tf-idf* measure to address possible content bias, but nonetheless a highly variable corpus

- What is driving this effect?
  - Little evidence offered for the L1-driven phonological hypothesis
Goals of present study

- Extend methodology to naturalistic speech data
- Use semantically controlled corpus to minimize variability in topic or register
- Explore classifier input in order to pinpoint the source(s) of the effect
The corpus
(Van Engen, Baese-Berk, Baker, Choi, Kim, and Bradlow, in press)

The Wildcat Corpus of Native- and Foreign-Accented English (from Northwestern University)

- Both scripted and spontaneous speech recordings
- Orthographically transcribed

- 24 native English speakers & 52 non-native English speakers
  - English (n=24), Korean (n=20), Mandarin Chinese (n=20),
    - Indian (n=2), Spanish (n=2), Turkish (n=2), Italian (n=1), Iranian (n=1),
    - Japanese (n=1), Macedonian (n=1), Russian (n=1), Thai (n=1)

- Designed in part to examine communication between talkers of different language backgrounds
Diapix task
(Van Engen, Baese-Berk, Baker, Choi, Kim, and Bradlow, in press)

<table>
<thead>
<tr>
<th>Changed Items</th>
<th>Version A</th>
<th>Version B</th>
<th>Missing Items</th>
<th>Version A</th>
<th>Version B</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat on pet shop sign</td>
<td>sheep on pet shop sign</td>
<td>no beehive</td>
<td>bee hive</td>
<td>no paw prints on door</td>
<td>no paw prints on door</td>
</tr>
<tr>
<td>pork chop sign</td>
<td>lamb chop sign</td>
<td>paw prints on door</td>
<td>Boss’s Booze</td>
<td>no sign</td>
<td></td>
</tr>
<tr>
<td>cheese soup</td>
<td>beef soup</td>
<td></td>
<td>just Pet Shop</td>
<td>Pete’s Pet Shop</td>
<td></td>
</tr>
<tr>
<td>woman has red shoes</td>
<td>woman has green shoes</td>
<td></td>
<td>no bench</td>
<td>bench</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>boy carrying box</td>
<td>boy not carrying box</td>
<td></td>
</tr>
</tbody>
</table>
# Subcorpus details

<table>
<thead>
<tr>
<th></th>
<th>English (n = 24)</th>
<th>Korean (n = 20)</th>
<th>Mandarin (n = 20)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word tokens</strong></td>
<td>15,617</td>
<td>17,253</td>
<td>19,168</td>
<td>52,038</td>
</tr>
<tr>
<td><strong>Word types</strong></td>
<td>981</td>
<td>927</td>
<td>915</td>
<td>1,461</td>
</tr>
<tr>
<td><strong>Word type/token ratio</strong></td>
<td>0.063</td>
<td>0.054</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td><strong>Unique character bigrams</strong></td>
<td>402</td>
<td>382</td>
<td>378</td>
<td></td>
</tr>
<tr>
<td><strong>Unique character trigrams</strong></td>
<td>2,141</td>
<td>2,006</td>
<td>1,982</td>
<td></td>
</tr>
</tbody>
</table>

Space = _  Apostrophe = ‘
k Nearest Neighbors (kNN)

- \( k = \) number of neighbors

- 1 speaker = 1 document = 1 vector
  - Multidimensional vectors of frequencies represent either: all words, all bigrams, or all trigrams

- Random 80% documents training, 20% testing
## Results

<table>
<thead>
<tr>
<th>k</th>
<th>Words</th>
<th>Bigrams</th>
<th>Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.2</td>
<td>69.5</td>
<td>69.2</td>
</tr>
<tr>
<td>4</td>
<td>53.8</td>
<td>61.5</td>
<td>76.9</td>
</tr>
<tr>
<td>8</td>
<td>69.2</td>
<td>61.5</td>
<td>69.2</td>
</tr>
</tbody>
</table>

(in percent correct)

Little decrease in accuracy after removing most frequent words
What is doing the classifying?

- Pick out n-grams that are:
  - maximally variant in frequency between language backgrounds
  - fairly frequent
What is doing the classifying?

- Look for possible phonological effects
  - Maybe English speakers use words with difficult consonant clusters that non-native speakers avoid?
So what *is* doing the classifying?

- A number of things...
Case 1: Single function word

N-gram significant because of one single function word

Other examples:
\texttt{ut}_\_ = ‘but’ and ‘about’
\texttt{_wi} and \texttt{1l}_\_ = ‘will’
Case 2: Single interjection

N-gram significant because of one single interjection or discourse marker

Other examples:
hm__ = ‘mhm’
yes = ‘yes’
no__ = ‘no’
Case 3: Single morpheme

N-gram significant because of one single morpheme

n’t
Combination of cases

Function and content words

Vocabulary items
Combination of cases

Content and function words

_\text{ca}_
Back to Tsur and Rappoport

- How generalizable are their results to speech?
  - Classifier performs well on orthographically transcribed speech

- Have we determined what is driving this effect?
  - Appears to be more lexical than phonological
Conclusions

▸ Can obtain successful classification using simple orthographic transcription
  – No phonetically or morphologically tagged corpus appears to be necessary

▸ Main action areas are morphosyntax and lexical semantics

▸ Classifier’s statistical power derived from collapsing across related cases
  – Trigrams do this best
Thank you:

Tyler Kendall

Bei Yu

Ann Bradlow

Language Dynamics Lab
at Northwestern University

Speech Communication Research Group
at Northwestern University
References


