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

AACL 2009 | University of Alberta | October 10

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

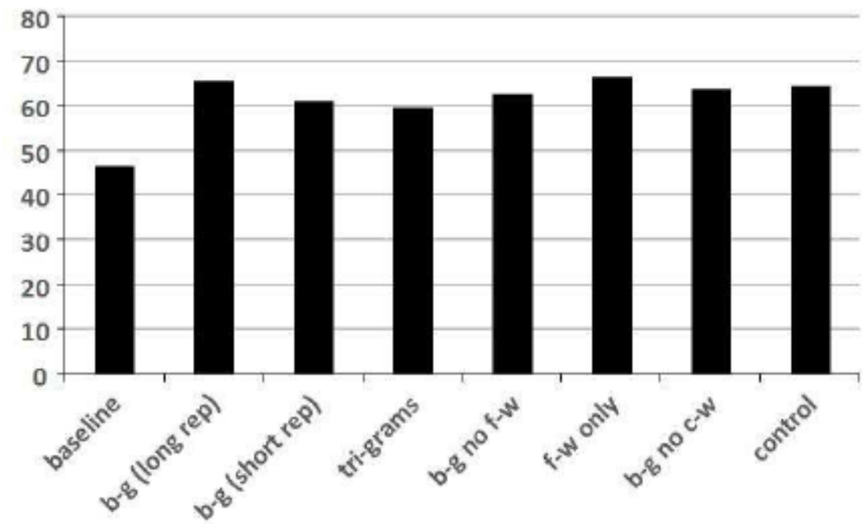
(Tsur and Rappoport, 2007)

- 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

Interpretation

(Tsur and Rappoport, 2007)

- 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)

Changed Items		Missing Items	
Version A	**Version B**	**Version A**	**Version B**
cat on pet shop sign	sheep on pet shop sign	no beehive	beehive
pork chop sign	lamb chop sign	paw prints on door	no paw prints on door
cheese soup	beef soup	Boss's Booze	no sign
woman has red shoes	woman has green shoes	just Pet Shop	Pete's Pet Shop
		no bench	bench
		boy carrying box	boy not carrying box

Subcorpus details

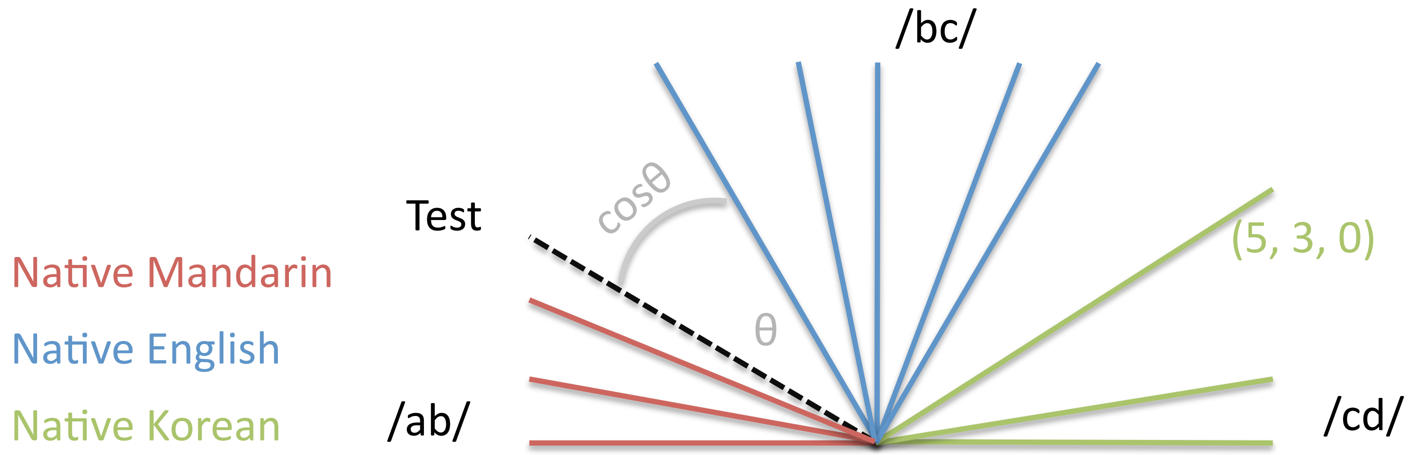
	English (n = 24)	Korean (n = 20)	Mandarin (n = 20)	Total
Word tokens	15,617	17,253	19,168	52,038
Word types	981	927	915	1,461
Word type/ token ratio	0.063	0.054	0.048	

Unique character bigrams	402	382	378
Unique character trigrams	2,141	2,006	1,982

Space = _ Apostrophe = ‘

Classifier

- 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

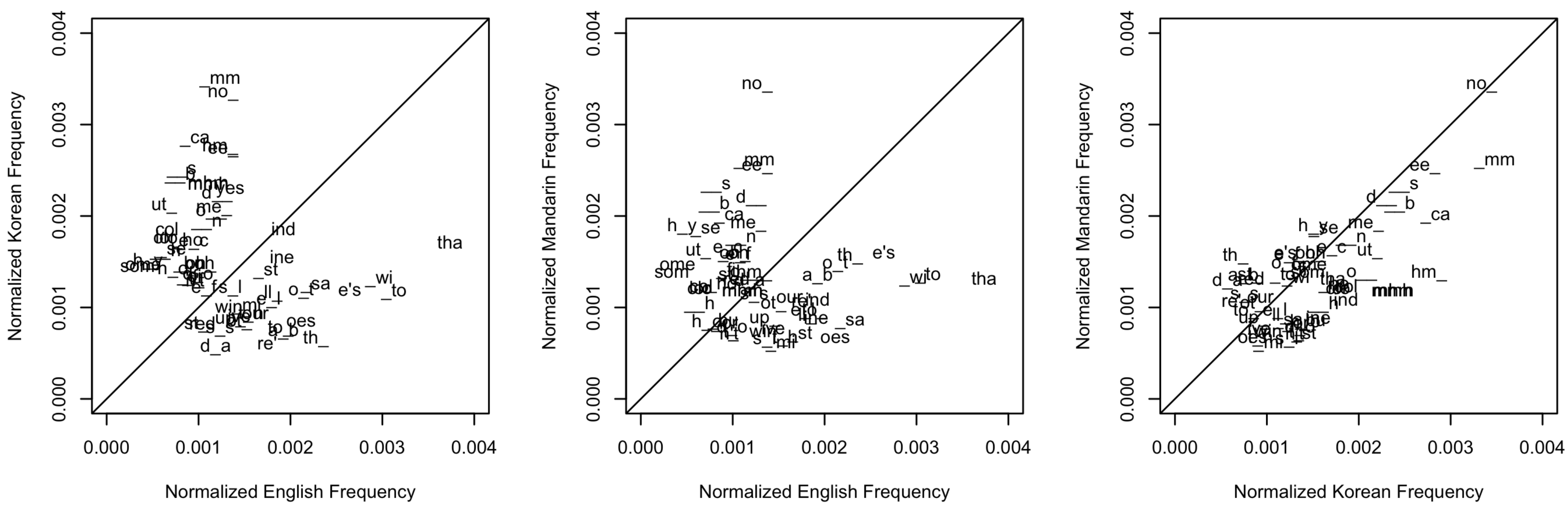
k	Words	Bigrams	Trigrams
1	69.2	69.5	69.2
4	53.8	61.5	76.9
8	69.2	61.5	69.2

(in percent correct)

Little decrease in accuracy after removing most frequent words

What is doing the classifying?

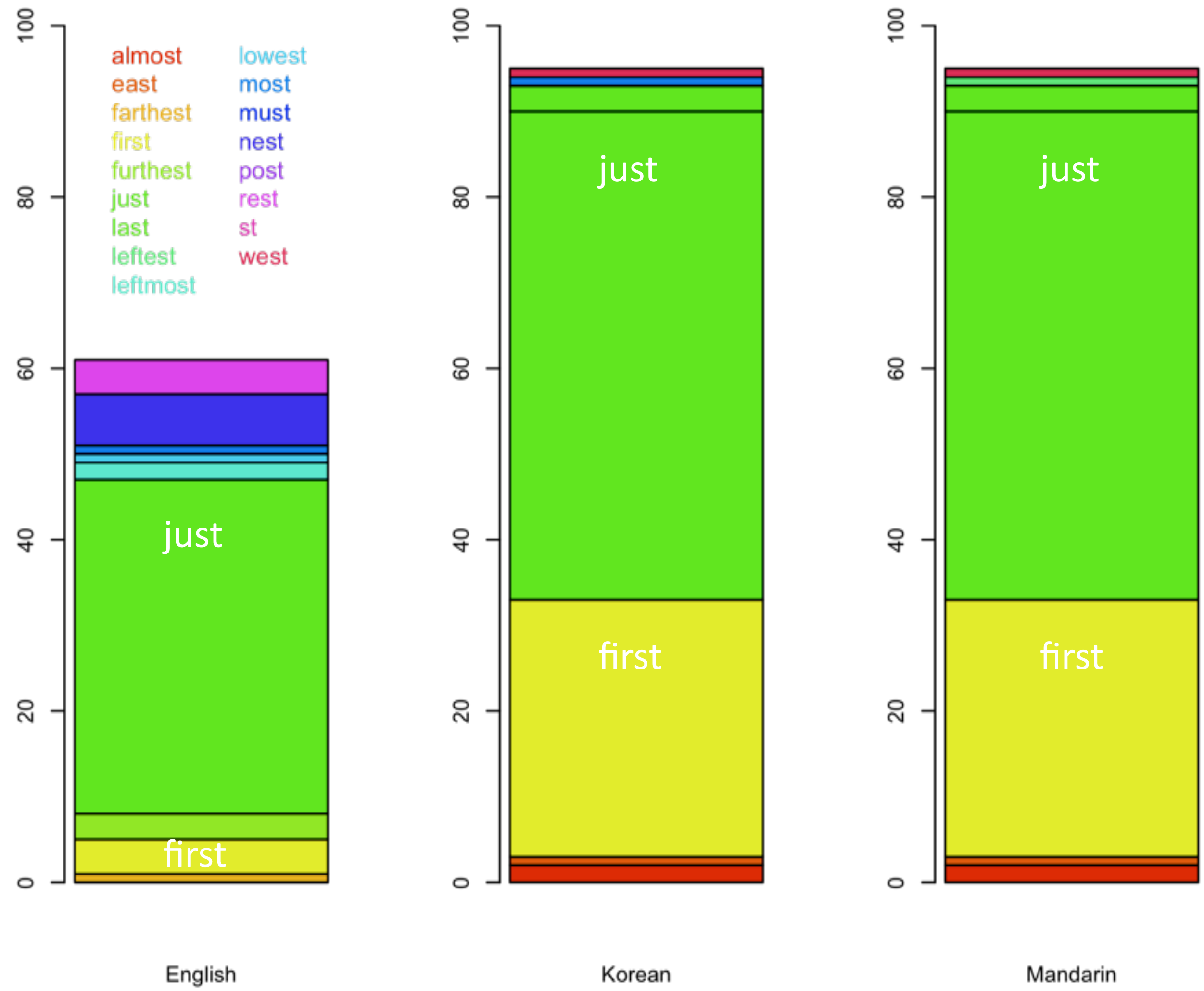
- Pick out *n*-grams that are:
  - maximally variant in frequency between language backgrounds
  - fairly frequent



What is doing the classifying?

- Look for possible phonological effects
  - Maybe English speakers use words with difficult consonant clusters that non-native speakers avoid?

st_

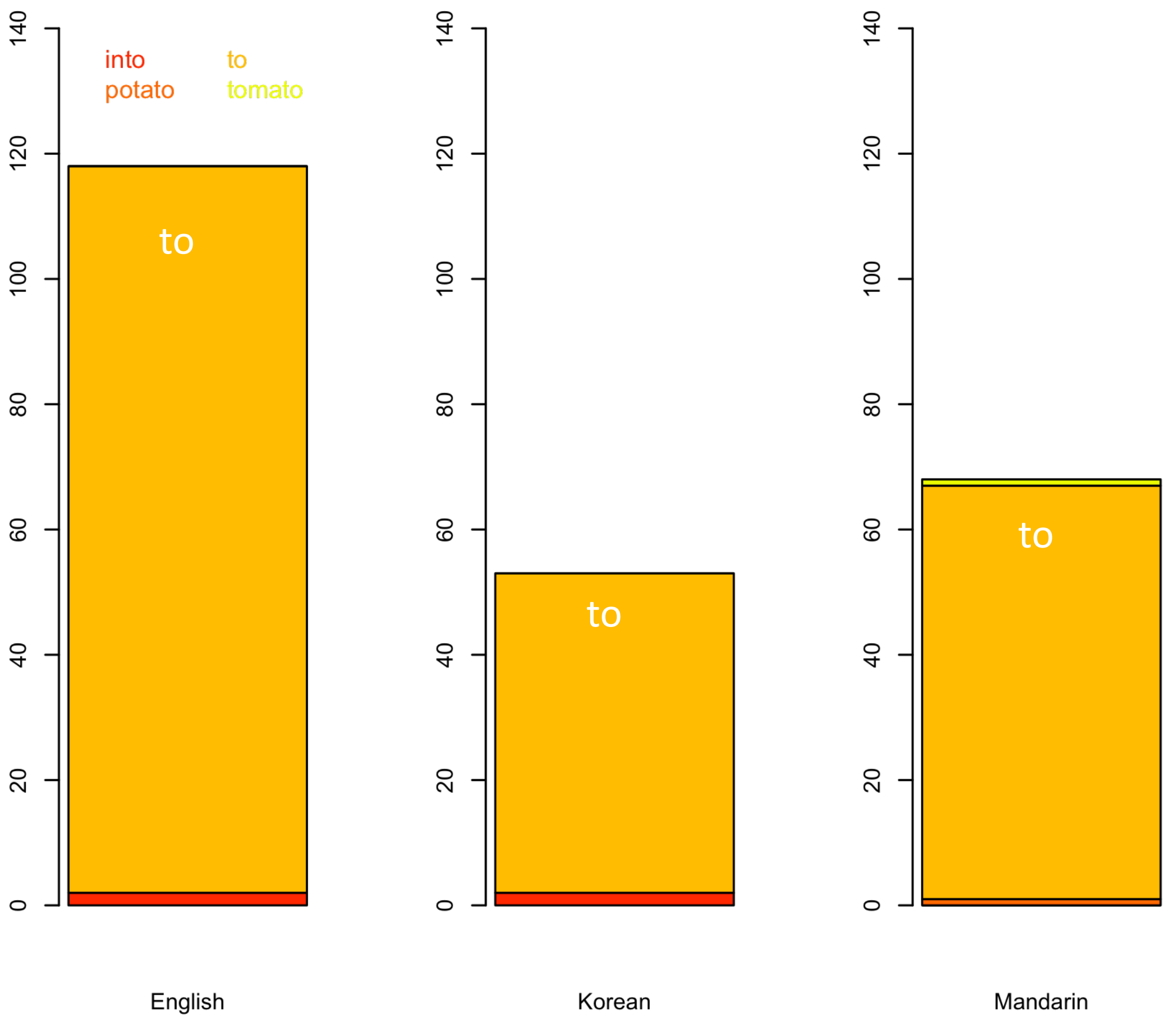


100
80
60
40
20
0
almost
east
farthest
first
furthest
just
last
leftest
leftmost
lowest
most
must
nest
post
rest
st
west
just
first
just
first
just
first
English
Korean
Mandarin

So what *is* doing the classifying?

- A number of things…

Case 1: Single function word



`to_`

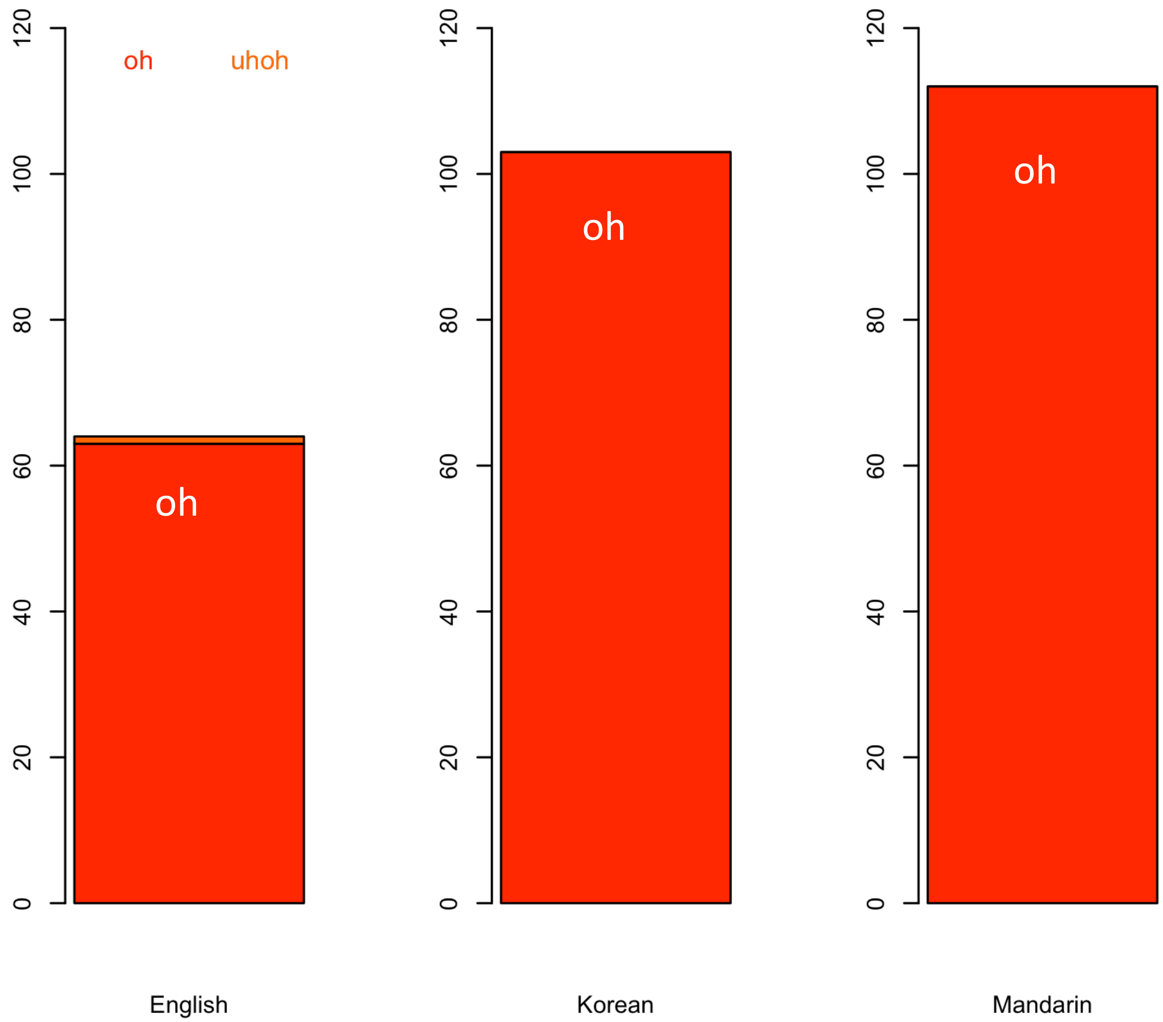
N-gram significant because of one single function word

Other examples:

`ut_` = ‘but’ and ‘about’

`_wi` and `ll_` = ‘will’

Case 2: Single interjection



`oh_`

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

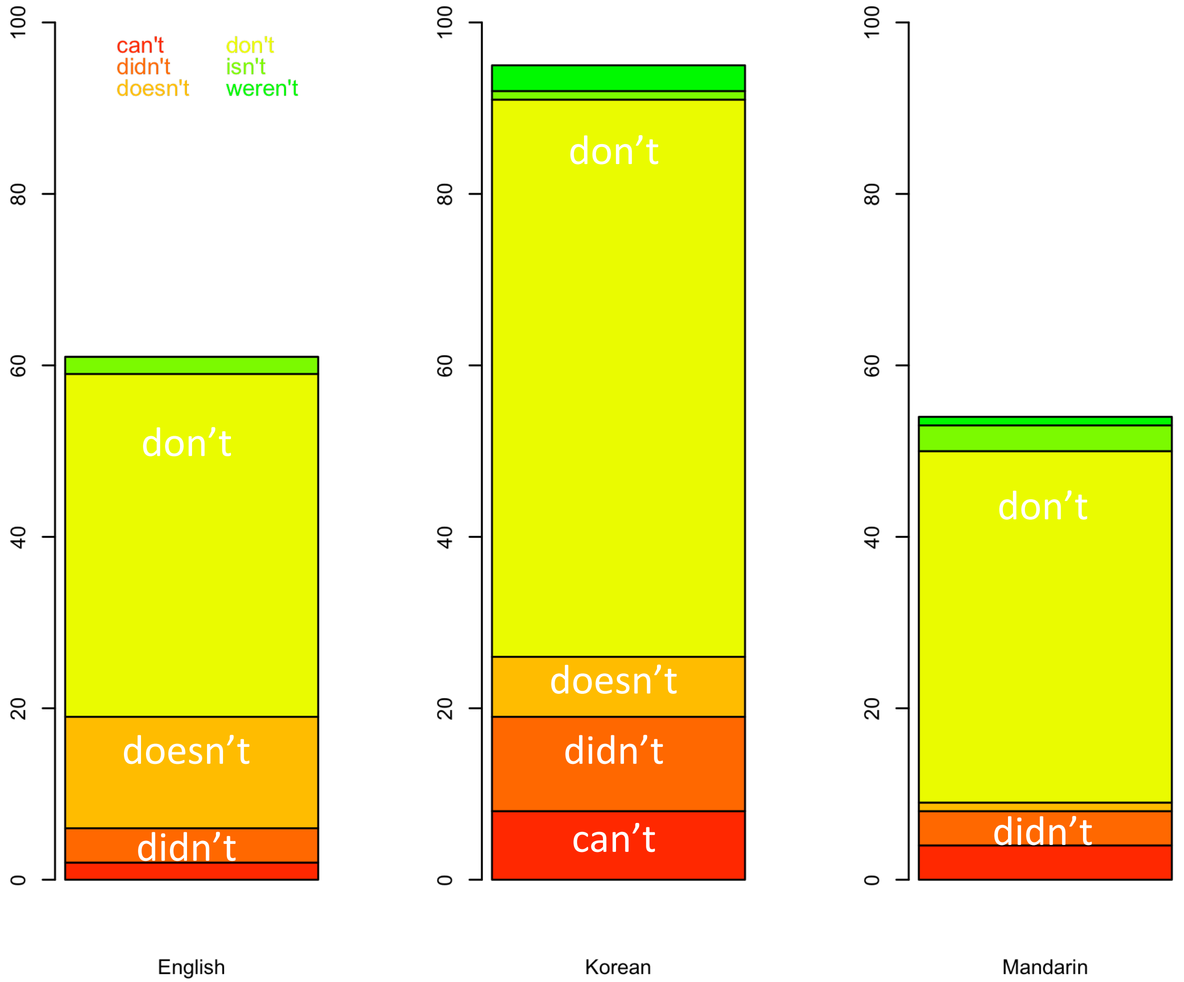
Other examples:

`hm_` = 'mhm'

`yes` = 'yes'

`no_` = 'no'

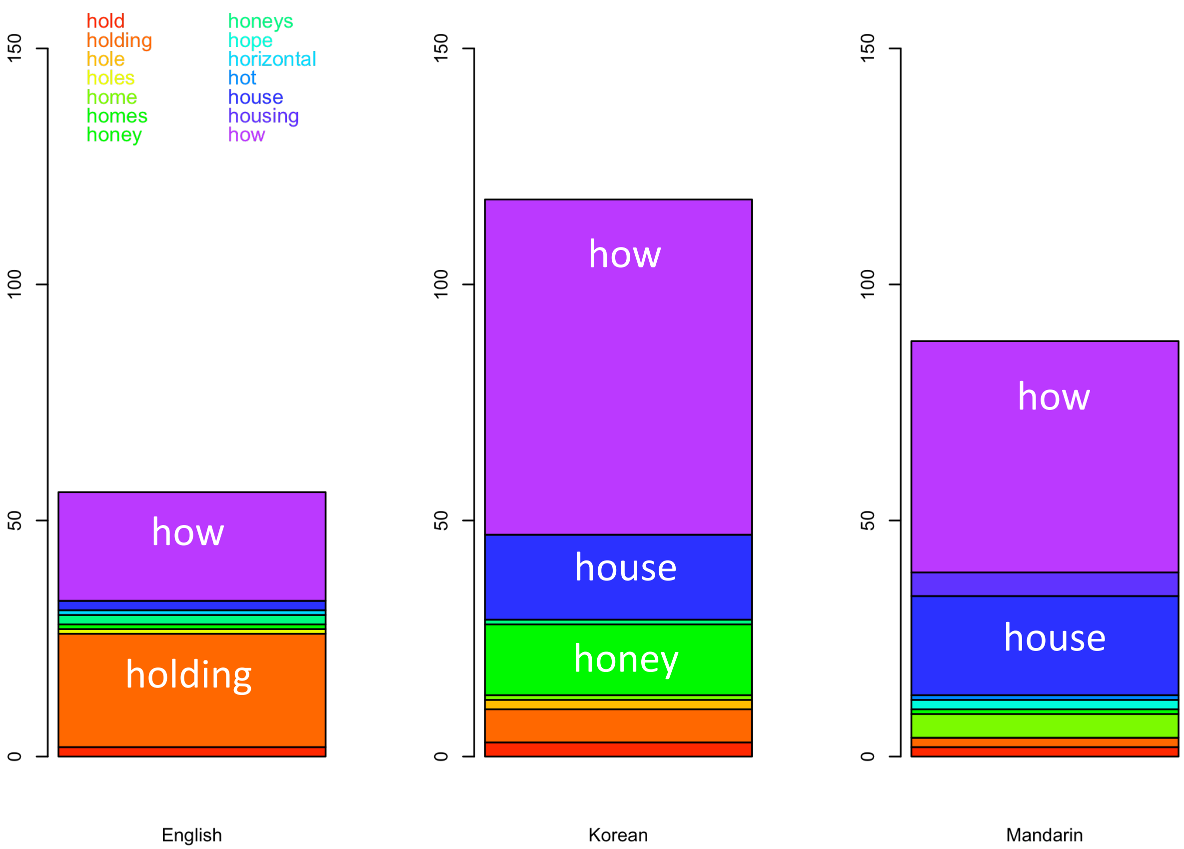
Case 3: Single morpheme



`n't`

N-gram significant because of one single morpheme

Combination of cases

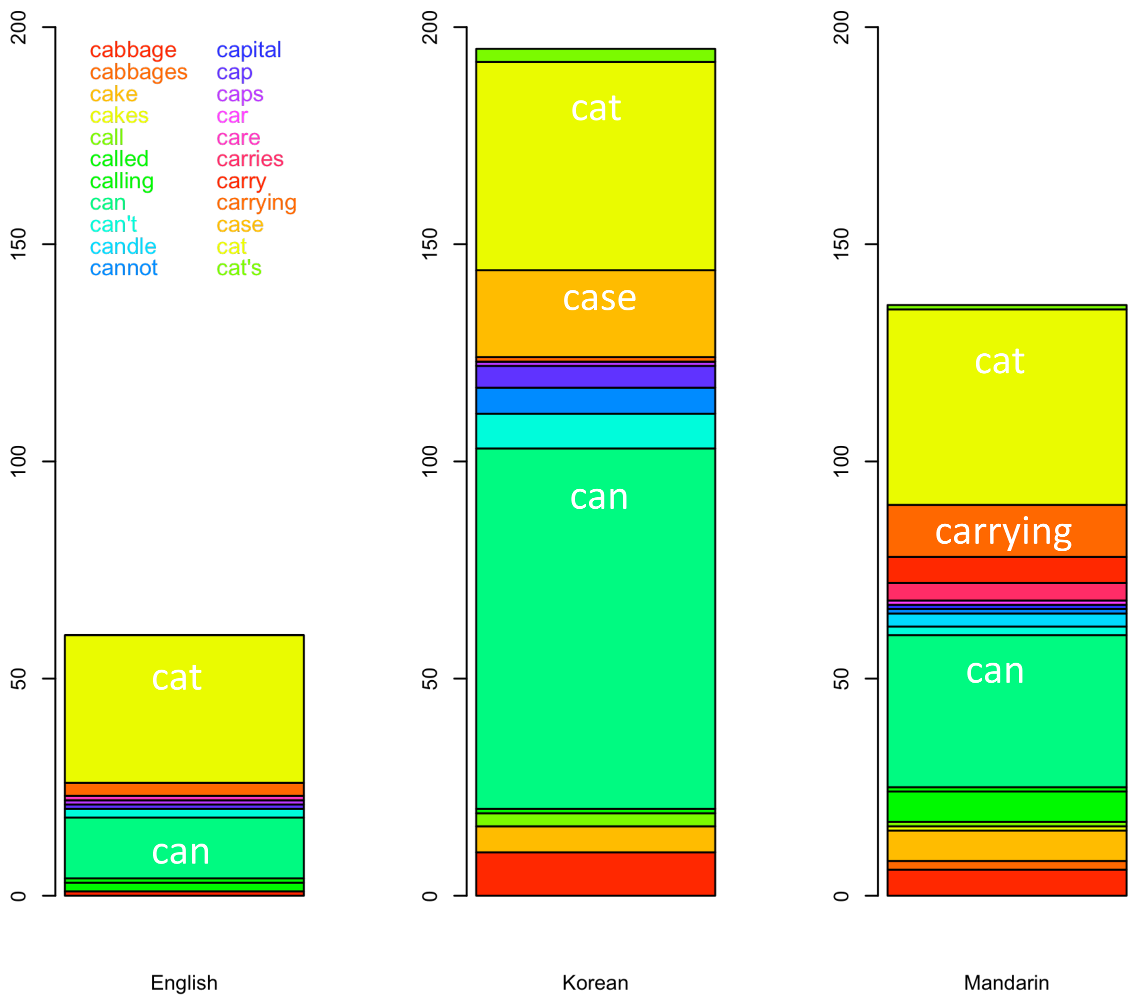


`_ho`

Function and content words

Vocabulary items

Combination of cases



_ca

Content and function words

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

Flege, J.E., 1987. The production of 'new' and 'similar' phones in a foreign language: evidence for the effect of equivalence classification. *J. Phonetics 15*, 47–65.

Flege, J.E., 1995. Second-language speech learning: theory, findings, and problems. In: Strange, W. (Ed.), *Speech Perception and Linguistic Experience, Issues in Crosslinguistic research*. York Press, Timonium, MD, 233–277.

Koppel M., J. Schler, and K. Zigdon K. 2005. *Automatically Determining an Anonymous Author's Native Language. In Intelligence and Security Informatics,* 209–217. Berlin / Heidelberg: Springer.

Mack, M., 2003. *The phonetic systems of bilinguals*. In: Banich, M.T., Mack, M. (Eds.), Mind, Brain, and Language: Multidisciplinary Perspectives. Lawrence Erlbaum Press, Mahwah, NJ.

Mosteller, F. and Wallace, D. 1964. *Inference and Disputed Authorship*, Addison – Wesley, Reading.

Tsur, O. and A. Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 6-16, Prague, Czech Republic, June 2007.

Van Engen, K., M. Baese-Berk, R. Baker, A. Choi, M. Kim, and A. Bradlow. In press. The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*.