Native and Non-native Spoken English

Katherine Kairis

Vienna-Oxford International Corpus of English (VOICE)

- Transcripts of face-to-face conversations
  - Includes part-of-speech-tagged versions of each conversation
- 1235 participants using English as common language
  - 49 L1s represented
  - Includes approximately 80 native English speakers

British National Corpus (BNC)

- Contains 100 million words (much larger than VOICE)
- Mixture of written and spoken texts by British English speakers
- Each word is tagged with its part of speech and lemma

Processing Data

- XML files – used BeautifulSoup
- Both VOICE and BNC required at least some modifications, but VOICE required much more.

```
participants['EDcon4_S1']
```

```
{'L1': ['pol'],
 'age': '17-24',
 'conversation': 'EDcon4.xml',
 'occupation': 'student',
 'role': 'participant',
 'sex': 'female'}
```

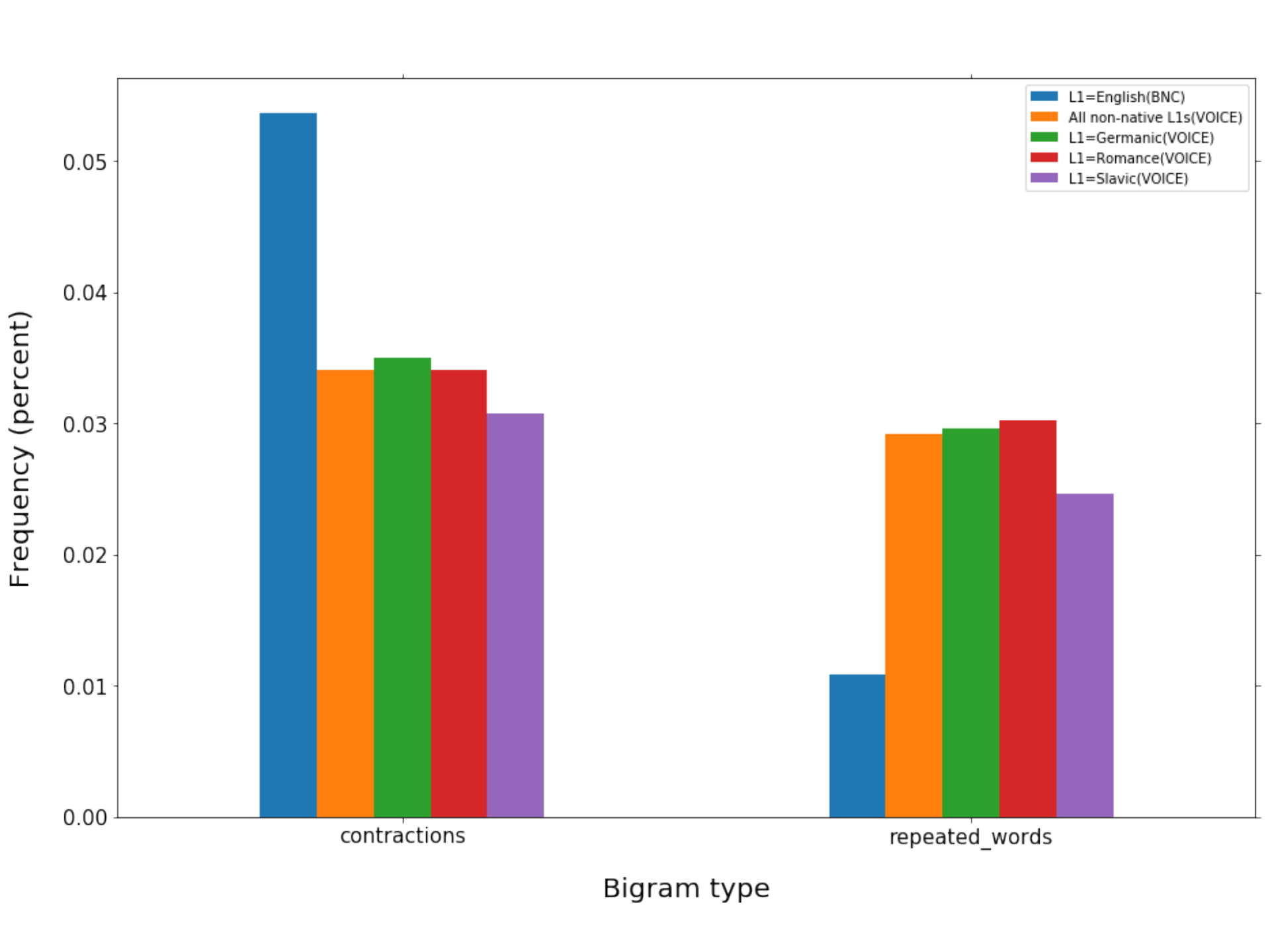
Processing Data

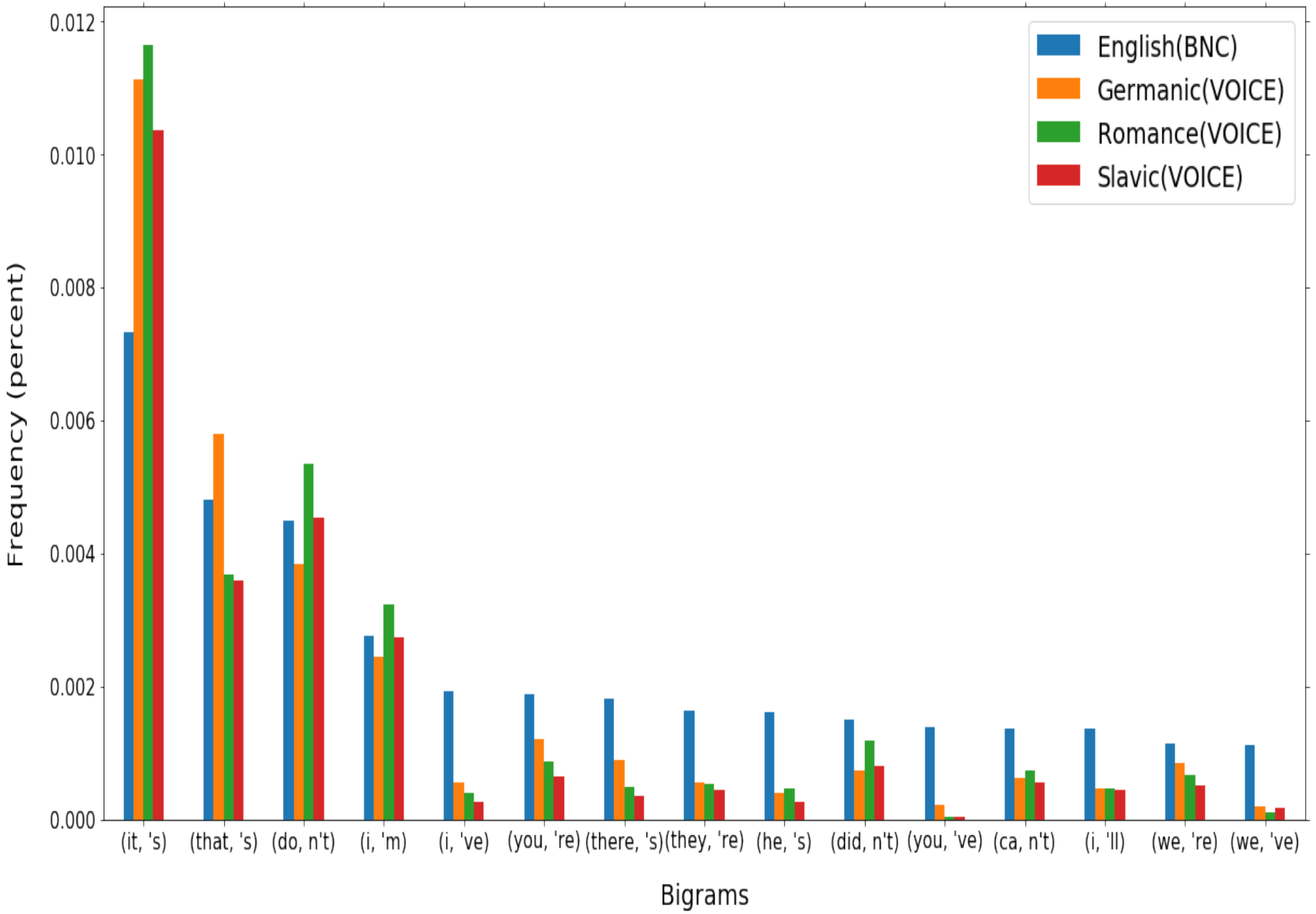
```
tokenized_conversations['EDcon4.xml']
```

```
{('EDcon4_u_1002', 'EDcon4_S2'): ['is', 'this', 'a', 'statement'],
 ('EDcon4_u_1003', 'EDcon4_S1'): ['no', 'you', "'re", 'overreacting'],
 ('EDcon4_u_1004', 'EDcon4_S2'): ['ah',
  'it',
  'means',
  'i',
  "'m",
  'clean',
  'that',
  'i',
  'wash',
  'my',
  'clothes',
  'and',
  'i',
  'wash',
  'myself',
  'also'],
 ('EDcon4_u_1005', 'EDcon4_S1'): ['i',
  'ca',
  "n't",
  'see',
  'him',
  'at',
  'town'],
 ('EDcon4_u_1006', 'EDcon4_S2'): ['what', 'you', "'re", 'what'],
 ('EDcon4_u_1007', 'EDcon4_S1'): ['[', 'first', 'name8', ']'],
 ('EDcon4_u_1008', 'EDcon4_S2'): ['yeah'],
```

Analysis

- Preliminary:
  - Bigrams
  - Stop words
  - Tokens
- Spent most time analyzing bigrams, since they showed some intriguing differences in the preliminary analysis.
- Compared 5 groups
  - Speakers in BNC
  - All participants in VOICE
  - Participants in VOICE with Germanic L1s
  - Participants in VOICE with Romance L1s
  - Participants in VOICE with Slavic L1s


L1=English(BNC)
All non-native L1s(VOICE)
L1=Germanic(VOICE)
L1=Romance(VOICE)
L1=Slavic(VOICE)
0.05
0.04
0.03
0.02
0.01
0.00
Frequency (percent)
contractions
repeated_words
Bigram type


English(BNC)
Germanic(VOICE)
Romance(VOICE)
Slavic(VOICE)
Frequency (percent)
0.012
0.010
0.008
0.006
0.004
0.002
0.000
(it, 's)
(that, 's)
(do, n't)
(i, 'm)
(i, 've)
(you, 're)
(there, 's)
(they, 're)
(he, 's)
(did, n't)
(you, 've)
(ca, n't)
(i, 'll)
(we, 're)
(we, 've)
Bigrams

Future plans

- Compare L1 groups for specific languages.
- Maybe try a classifier with some of the features that I discover.