Native and non-native speakers’ idiom production: What can read speech tell us?

Jing Liu\textsuperscript{1} & Helmer strik\textsuperscript{1,2,3}

\textsuperscript{1}Center for language studies, \textsuperscript{2}Centre for Language and Speech Technology, \textsuperscript{3}Donders Institute for Brain, Cognition and Behaviour

Radboud University Nijmegen, the Netherlands

MWE 2022 Workshop
06/25/2022
Idiom production

• Holistic representation in mental lexicon
Idiom production

• Holistic representation in mental lexicon
  - Comprehension

Eye-tracking (Jiang et al., 2020; Siyanova-Chanturia et al., 2011)
Self-paced reading paradigm (Conklin & Schmitt, 2008)
Idiom production

- **Holistic representation in mental lexicon**
  - Comprehension: Shorter fixation/reading time (e.g. Jiang et al., 2020)
  - Production: Superlemma Theory (Kuiper et al., 2007)

- Unitary idiomatic concept
Idiom production

• Holistic representation in mental lexicon
  - Comprehension: Shorter fixation/reading time (Jiang et al., 2020)
  - Production: Superlemma Theory (Kuiper et al., 2007)

- Unitary idiomatic concept
  -> holistic phonological encoding
  -> higher co-occurrences

Less hesitation, pauses and dysfluencies
Idiom production

• Holistic representation in mental lexicon
  - Comprehension
    Eye-tracking (Jiang et al., 2020; Siyanova-Chanturia et al., 2011)
    Self-paced reading paradigm (Conklin & Schmitt, 2008)
  - Production
    Superlemma Theory (Kuiper et al., 2007)
    Longer formulation time (Goldberg, 2017)

Q: articulation level?
Idiom production

- **Holistic representation in mental lexicon**
  - Comprehension
    Eye-tracking (Jiang et al., 2020; Siyanova-Chanturia et al., 2011)
    Self-paced reading paradigm (Conklin & Schmitt, 2008)
  - Production
    Superlemma Theory (Kuiper et al., 2007)
    Longer formulation time (Goldberg, 2017)
  - Influence of language proficiency (Zheng et al., 2021)

Q: L2 speakers?
Research design

• Focus on articulation level

Read speech
Research design

• Focus on articulation level

• Bottom-up investigation
  - Include all possible acoustic features

Multidimensional acoustic embedding (Beguš & Zhou, 2022)
Research questions

Q1: What are the differences of idiomatic and non-idiomatic speech produced by native speakers?

Q2: What are the differences of idiomatic and non-idiomatic speech produced by L2 speakers?
Methodology

• Idiom selection

44 idioms were selected from 393 idioms based on:

<table>
<thead>
<tr>
<th>Property</th>
<th>Explanation</th>
<th>Native</th>
<th>L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Relative degree to which participants have come across an idiom in speech or in print</td>
<td>0.94</td>
<td>0.88</td>
</tr>
<tr>
<td>Familiarity</td>
<td>How well speakers know the meaning of an idiom</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Transparency</td>
<td>The extent to which the original metaphorical meaning can be deduced from its literal analysis</td>
<td>0.94</td>
<td>0.90</td>
</tr>
<tr>
<td>Imageability</td>
<td>The extent to which an idiom can evoke an image</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>Objective knowledge</td>
<td>Obtained from a multiple-choice test of meaning recognition</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Usage</td>
<td>Frequency subjects indicate having used an idiom</td>
<td>0.91</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Methodology

• Data collection

L2 speech data (ISLA corpus)
- Produced by 22 German L1 learners of Dutch L2 (mean age = 24.76; SD = 3.46)
- Each idiom is embedded in a sentence
- 1413 recordings altogether

Native speech material (Spoken Dutch Corpus: CGN)
- Corresponding 44 idioms produced by native Dutch speakers
- Each idiom is embedded in the sentence
- 175 recordings altogether
Data analysis

Step 1: Data pre-processing

- Forced alignment
-> Get temporal information for each word
Automatic: CLARIAH
Manual checking

- Idiom extraction
Automatic: Lemma-based matching
Manual checking
Step 2: Feature extraction on the word level
- Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS: Eyben et al., 2016)
88 Frequency related, energy related, spectral and temporal parameters

- Praat features
17 speech rate, articulation rate, intensity, formants, center of gravity related features

-> 105 features altogether
Step 2: Feature extraction on the word level
- Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS: Eyben et al., 2016)
- Praat features
-> 105 features altogether

Characteristic for some types of atypical read speech
1) Read Dutch COPD speech (Van Bemmel et al., 2021)
2) Dysarthric speech (Wei et al., 2021)
3) Non-native speech (Bosland, 2022)
### Data analysis

#### Step 3: Feature ranking

- Outlier detection
- Recursive Feature Elimination (RFE: Guyon et al., 2002)

#### Data:

\[ D = \{X, L\} \]  // a training data set with \( n \) number of features

\[ X' \]  // predefined initial feature subset (\( X' \subset X \) or \( X' = \{\} \))

\( \theta \)  // a stopping criterion

**Result:** \( X'_{\text{opt}} \)  // an optimal subset

```
begin

initialize:
\[ X_{\text{opt}} = X' \]
\[ \phi_{\text{opt}} = E(X', A) \]  // evaluate \( X' \) by using an mining algorithm \( A \)

while (\( \theta \) not reached) do

\[ X_g = \text{generate}(X) \]  // subset generation for evaluation

\[ \phi = E(X_g, A) \]  // \( X_g \) current subset evaluation by \( A \)

if (\( \phi > \phi_{\text{opt}} \)) then

\[ \phi_{\text{opt}} = \phi \]

\[ X_{\text{opt}} = X_g \]

return \( X'_{\text{opt}} \)
```

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>...</th>
<th>Feature n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
<tr>
<td>Word 2</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word n</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature # 1</th>
<th>Feature #2</th>
<th>...</th>
<th>Feature #n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
<tr>
<td>Word 2</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word n</td>
<td>XXX</td>
<td>XXX</td>
<td>XXX</td>
</tr>
</tbody>
</table>
Step 4: Statistical analysis

- T-tests

1.) Individual effects of features
2.) Rank all the features based on significance and effect size

- Support Vector Machine Model (SVM)
Select the most characteristic combination of features based on performance

- Principal Component Analysis (PCA)
Reduction of overlapping features
Results

Idioms v.s. Non-idioms produced by native speakers

T-tests: 45 significant features; SVM: 34 features
-> 13 overlapping features

Table 1
Feature categories after PCA

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th>%variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness</td>
<td>Loudness percentile 20.0 - 50.0; intensity mean; loudness std</td>
<td>23.965</td>
</tr>
<tr>
<td>Spectral flux</td>
<td>Spectral Flux std; Spectral Flux normalized std</td>
<td>14.381</td>
</tr>
<tr>
<td>Speaking rate</td>
<td>Articulation rate(syllable/sec); Speech rate(syllable/sec)</td>
<td>10.956</td>
</tr>
<tr>
<td>Formant</td>
<td>F3; F2bandwidth normalized std; F2 mean amplitude</td>
<td>8.416</td>
</tr>
<tr>
<td>F0</td>
<td>F0 semitone percentile20.0; pitch std; F0semitone Falling Slope</td>
<td>7.348</td>
</tr>
</tbody>
</table>
## Results

### Idioms v.s. Non-idioms produced by L2 speakers

T-tests: 45 significant features; SVM: 26 features

-> 9 overlapping features

**Table 2**

*Feature categories after PCA*

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th>%variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness and spectral flux</td>
<td>Loudness percentile20-80; intensity mean; intensity min equivalent Sound Level dB; spectral Flux mean</td>
<td>41.746</td>
</tr>
<tr>
<td>Formant</td>
<td>F3amplitude mean; F2amplitude mean; F1amplitude mean</td>
<td>22.108</td>
</tr>
<tr>
<td>F0</td>
<td>F0 semitone percentile20.0</td>
<td>10.258</td>
</tr>
</tbody>
</table>
Conclusion

• Speaking rate

Native speakers: *shorter* speech and articulation rate in idiomatic expressions
L2 speakers: no significant difference in speech and articulation rate

-> The superlemma representation may be influenced by language proficiency
Conclusion

• **Speaking rate**

Native speakers: *shorter* speech and articulation rate in idiomatic expressions
L2 speakers: no significant difference in speech and articulation rate
-> The superlemma representation may be influenced by language proficiency

• **Loudness**

Higher loudness in idiomatic expressions for both native and L2 speakers
-> Higher confidence in the more frequent fixed collocations (Cucchiarini, 2000)
-> Syntactic structure of the idioms (most are verb phrases)
Limitations & future research

• **Data differ between idioms and non-idioms**
  Formant as one of the characteristic features (different words)
  Loudness as one of the characteristic features (different syntactic functions)
  -> Next step: Collect similarly structured non-idiom materials

• **Limited (native) speech data may lead to overfitting**
  -> Next step: Collect more (native) speech data
References


References


Acknowledgements

• Special thanks to ISLA Team who inspired me to explore the magic of idioms and their support in formulating hypotheses.

• Special thanks to AASPA group (Loes van Bemmel, Xing Wei etc.) whose help was invaluable in developing the scripts.
Thanks for your attention!

For more information:

@JLiu_Compuling  @Jing-L97