Beyond two-word sequences: Assessing phraseological development in L2 texts using directional association measures

Alvin Cheng-Hsien Chen

National Taiwan Normal University, Taipei, Taiwan

Presented at TaLC 2018, 19 July, Cambridge, UK
Phraseology

• A general linguistic phenomenon that words often co-occur as bundles of variable lengths
• From an intuition-based top-down approach to distribution-based bottom-up approach
  - Corpora data
    • Learner corpora, Representative L1 corpora
  - Advanced applications
    • *Wordsmith, AntConc, Lextutor, TAALES, LancBox*, and more
This Study

• Previous studies on L2 Phraseology
  – Type/token frequencies (Cortes, 2004; Leńko-Szymańska, 2014; Qin, 2014)
  – degree of formulaicity (Bestgen, 2017; Bestgen & Granger, 2014; Durrant & Schmitt, 2009)

• Research Questions
  – Does the development of L2 phraseological competence differ in terms of words co-selection in different directions?
  – How is this development mediated by the length of the bundles considered?
Data

• Learner Corpus
  – International Corpus Network of Asian Learners of English 2.0 (Ishikawa, 2013)
  – 5,200 essays
  – Learners were annotated with proficiency levels based on CEFR labels (A2, B1_1, B1_2, B2+)

• Reference Native Corpus
  – COCA (2012)
  – BNC XML version
Estimating phraseological metrics using COCA

- All texts were segmented into chunks by taking non-word tokens as delimiters.
- N-grams from N=2 to N=5 were extracted based on chunks.
- N-grams with numbers were removed.
- All n-grams were normalized into lower-case letters.
- Cut-offs (Frequency and Dispersion):
  - Frequency > 10 for bigrams and trigrams; frequency > 5 for 4-grams and 5-grams.
  - Range >= 5.
Formulaicity vs. Delta P

• The metric of Delta P has now been well-received as an effective statistical metric for measuring directional associations of:
  – linguistic co-occurrences (Ellis & Ogden, 2017; Gablasova et al., 2017; Gries, 2013; Wahl, 2015)
  – behavioral data (Ellis & Ferreira-Junior, 2009)
• If a multiword sequence is formulaic, one word in the bundle should be somewhat conditioned by its co-texts
  – [and he → ...] (was, said) vs. [so he → ...](could, can)
  – [... ← was told] (I)
Delta P

**Contingency Table for Delta P Computation**

<table>
<thead>
<tr>
<th></th>
<th>Outcome</th>
<th>¬Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cue</strong></td>
<td>(O_{11})</td>
<td>(O_{12})</td>
</tr>
<tr>
<td>¬<strong>Cue</strong></td>
<td>(O_{21})</td>
<td>(O_{22})</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>(C_1)</td>
<td>(C_2)</td>
</tr>
</tbody>
</table>

\(R_1\) \(R_2\) \(N\)

*Notes. \(O\) refers to the observed frequencies of each cell; \(R\) refers to the sums of the rows; \(C\) refers to the sums of the columns.*

- The conditional probability of an **outcome** given the **cue** minus the conditional probability of an **outcome** in the absence of the **cue**
- A multiword sequence
  - **Outcome** as the choice of the last/first word
  - **Cue** as the preceding/following co-texts (Directionality) of variable sizes (Order)

\[
P(\text{outcome}|\text{cue}) - P(\text{outcome} | ¬\text{cue}) = \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2}
\]
Contingency Learning

• How likely a native speaker would choose the word:
  — Given the same preceding or subsequent co-texts

• We can manipulate the positions of the co-texts and the sizes of the co-texts to see how directionality and lengths may play a role
Number of Parameters in COCA

<table>
<thead>
<tr>
<th>N-gram Type</th>
<th>BNC</th>
<th>COCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2gram</td>
<td>65,1278</td>
<td>2,334,463</td>
</tr>
<tr>
<td>3gram</td>
<td>68,9826</td>
<td>3,295,239</td>
</tr>
<tr>
<td>4gram</td>
<td>74,6042</td>
<td>4,624,772</td>
</tr>
<tr>
<td>5gram</td>
<td>21,9841</td>
<td>1,843,582</td>
</tr>
</tbody>
</table>

*Notes.* These numbers refer to the total number of n-gram types that passed the distributional cut-offs described in Section 3.2.
Normalization of COCA-based DPs

• As DP increases by its nature with the lengths of the cue, all DP values were normalized into z-scores on an ngram basis.
Profiling L2 Texts

L2 Text → Ngrams (N = 2 ~ 5)

Delta P Assignment → 8 Means Scores (4 Lengths in 2 Directions)

19 July, 2018

alvinworks@gmail.com
Statistical Analysis

• Mixed-design Multilevel Linear model
  – Analyze the variation of the DP mean scores of each L2 text in relation to:
    – Between-subject factor:
      • Proficiency Level: A2/B1_1/B1_2/B2+(LEVEL)
    – Within-subject factors:
      • Directions of the DP: FWD/BWD (DIRECTION)
      • Cues Lengths (i.e., ngram order): 2/3/4/5 (ORDER)
  – All high-order interactions
Statistical Results

- **LEVEL × DIRECTION × ORDER** ($p < 0.01$)
  - Main Effect
  - 2-way: **LEVEL × DIRECTION**
  - 3-way
Main Effect (I): LEVEL

- DP increases with L2 proficiency (LEVEL), \( \chi^2(3) = 150.89, p < 0.01, r = 0.12 \).

- Disregarding DIRECTION and ORDER, learners develop this native-like intuition in word co-selection.
Main Effect (II): DIRECTION

- **BWD DP is significantly larger in L2 texts**, \( \chi^2(1) = 930.95, p < 0.01, r = 0.36. \)

- **Learners significantly use more bundles with stronger BWD DPs**
Main Effect (III): Order

- The higher the order of the ngram, the lower the DP, $\chi^2(3) = 21679.3$, $p < 0.01$, $r = 0.52$
- Longer sequences, if used by L2 learners, tend to have lower DP values (less formalic?)
2-way: **Level x Direction**

- Learners develop more significantly on the FWD DPs from the early stage of proficiency onwards (bottom line).
- BWD DPs are on average higher in L2 texts than FWD DPs at all proficiency levels (top line).
- Development in BWD DP is more pronounced in advanced levels.
3-way Interaction

- As the cue length of the DP (ORDER) grows, the increasing trends of FWD and BWD DP start to converge
- Directional differences in L2 development are lessened in longer bundles
Implications

• High BWD DP on average
  — English is a head-first right-branching language
  — A higher backward transitional probability of a bigram significantly correlates with higher phrase cohesiveness between two words, i.e., tighter constituents in English (Onnis and Thiessen, 2013)

• High BWD DP suggests that learners pick up this general right-branching structure in English

• The increase in backward DP as an indicator of development of phraseology in-between tighter constituents, thus leading to a higher level of phrasal cohesiveness in writing.
Conclusions

• Formulaicity increases with learner proficiency

• When directionality is considered, learners show a steadier increase in forward phraseological competence across all proficiency levels

• Backward competence, however, improves more markedly at advanced levels

• Longer bundles are probably more difficult to master for learners
Thank you
Q & A