Corpus approaches to collocation: Challenges and opportunities

Dr. Vaclav Brezina & Dr. Dana Gablasova
Lancaster University
In essence, collocation is a phenomenon concerned with repeated co-occurrence of words in texts. There is something profoundly simple, yet exceptionally insightful about the immediate space that words share with each other in texts (Brezina 2018:59).
Where to start?
What do we know about collocation?

1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.
What do we know about collocations?

1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.
What do we know about collocations?

1. There is no consensus about the nature of the phenomenon.
2. There is no consensus about how to identify collocations.
3. There is no consensus about the terminology.
4. It is important.
Did you mean: "collocations" OR "formulaic language" OR "bigrams" OR "lexical bundles" OR "phraseology" OR "fixed expressions" OR "multi-word expressions" OR "prefabricated chunks" OR "prefabricated language" OR "idioms"

[BOOK] Formulaic language and the lexicon
A Wray - 2005 - books.google.com
Co-location as a starting point


Source: Wray (2002: 9)
Co-location as a starting point

amalgams – automatic – chunks – clichés – co-ordinate constructions –
colloctions – complex lexemes – composites – conventionalized
formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen
simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword
speech – precoded conventionalized routines – prefabricated routines and patterns – ready-
made expressions – ready-made utterances – recurring utterances – rote – routine formulae
– schemata – semipreconstructed phrases that constitute single choices – sentence builders
– set phrases – stable and familiar expressions with specialized subsenses – stereotyped
phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech –
unanalyzed multiword chunks – units

Source: Wray (2002: 9)
Co-location as a starting point (cont.)

- Low-inference category.
- Pre-theoretical.
- Data-driven.
- Close to textual reality.
Reasons for co-location

1. Semantic unit (*carbon monoxide, global warming, okey dokey*).
2. Lexico-grammar (*of the, difference between*).
3. Register preference (*large difference, administer a test, fucking stupid*).
4. Sociolinguistic choice (*sick movie, I believe*).
5. Discourse prosody (*illegal immigration, frail elderly.*)
Reasons for *studying* collocation

1. Language description (grammar, lexis, pragmatics etc.).
2. Discourse analysis (social, historical etc. meanings).
3. Language acquisition (L1 and L2).
4. Language pedagogy.
5. Language testing.
Corpus linguistics
Collocations

My love is like a red, red rose that’s newly sprung in June: My love is like the melody that’s sweetly played in tune. As fair art thou, my bonnie lass, so deep in love am I: And I will love thee still, my dear, till a’ the seas gang dry. Till a’ the seas gang dry, my dear, and the rocks melt wi’ the sun: And I will love thee still, my dear, while the sands o’ life shall run. And fare thee weel, my only love, and fare thee weel a while! And I will come again, my love, thou’ it were ten thousand mile.

(Robert Burns, “A Red, Red Rose”)

collocation window (span): 1L 1R
Random baseline model

‘my love’ ... 3

My love is like a red, red rose that’s newly sprung in June: My love is like the melody that’s sweetly played in tune. As fair as a stranger, a stranger will love thee still, my dear, till a’ the seas gallons rocks melt wi’ the sun: And I will love thee still, my dear, while the sands o’ life shall run. And fare thee weel, my only love, and fare thee weel a while! And I will come again, my love, thou’ it were ten thousand mile.

‘my love’ ... 1

fare art And like red, sweetly in love love, And gang wi’ played like dear, life shall rocks sprung the Till deep my my And still, weel, again, ten the the while! is till And As I: a’ only come were sands sun: My bonnie My red is a run. my love thee thou’mi am rose love dear, that’s love newly love fare love, will o’ so dry. fair thee will that’s in while June: my seas tune. mile. thousand weel dear,
Association measures

\[ \log_2 \frac{O_{11}^3}{E_{11}} \]

- Log likelihood

\[ \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \]

- Z score

\[ \log_2 \frac{O_{11}^2}{E_{11}} \]

- Log ratio

\[ \frac{O_{11} \times R_2}{E_{11} \times R_{1\text{cor}}} \]

- Cohen's d

\[ \frac{O_{11}}{R_1} - \frac{O_{21}}{R_2} \]
Association measures (cont.)
Association measures (cont.)
Dimensions of collocation

1. Frequency of co-occurrence
   - Make a decision vs pay obeisance

2. Exclusivity
   - love affair ↔ love you
   - guinea pig, carbon monoxide

3. Directionality
   - extenuating → circumstances; circumstances → extenuating?
   - love → you; you → ?

4. Distance (span)

5. Connectivity (collocation networks)
In essence, collocation is a phenomenon concerned with repeated co-occurrence of words in texts. There is something profoundly simple, yet exceptionally insightful about the immediate space that words share with each other in texts.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
<th>Significance</th>
<th>MINIMUM SENSITIVITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU:</td>
<td>13.672</td>
<td>LL: 1,526.896 (p &lt; 0.0001)</td>
<td>0.000</td>
</tr>
<tr>
<td>MI:</td>
<td>3.773</td>
<td>Z-score: 70.570</td>
<td>0.000</td>
</tr>
<tr>
<td>MI2:</td>
<td>12.501</td>
<td>T-score: 19.085</td>
<td>0.000</td>
</tr>
<tr>
<td>MI3:</td>
<td>21.229</td>
<td>DICE: 0.000</td>
<td>0.000; 0.2191</td>
</tr>
</tbody>
</table>

BNC: 424 occurrences (3.78 per million words)
Visualizing collocations
## Traditional form of display

<table>
<thead>
<tr>
<th>Collocate</th>
<th>MI-score</th>
<th>Freq (coll.)</th>
<th>Freq (corpus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>affair</td>
<td>8.86</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>fell</td>
<td>8.52</td>
<td>14</td>
<td>131</td>
</tr>
<tr>
<td>falling</td>
<td>8.52</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>fallen</td>
<td>8.37</td>
<td>5</td>
<td>52</td>
</tr>
<tr>
<td>me</td>
<td>5.57</td>
<td>23</td>
<td>1667</td>
</tr>
<tr>
<td>i'm</td>
<td>5.30</td>
<td>5</td>
<td>437</td>
</tr>
<tr>
<td>life</td>
<td>5.12</td>
<td>8</td>
<td>791</td>
</tr>
</tbody>
</table>
Collocation graph
Collocation networks

C2
C1
N4
C1
N3
C2
N2
C2
C3
C5
C4
C3
node
Everyday

#LancsBox
Parameters
Span

5L, 5R: 36 collocates

1L, 1R: 8 collocates
Figure 1 Top ten collocations of *make* for frequency and three AMs using L0, R2 windows in BE06 corpus.
Threshold

Low threshold

High threshold
### CPN (Brezina et al. 2015)

<table>
<thead>
<tr>
<th>Statistic ID</th>
<th>Statistic name</th>
<th>Statistic cut-off value</th>
<th>L and R span</th>
<th>Minimum collocate freq. (C)</th>
<th>Minimum collocation freq. (NC)</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>4b</td>
<td>MI2</td>
<td>3</td>
<td>L5-R5</td>
<td>5</td>
<td>1</td>
<td>function words removed</td>
</tr>
</tbody>
</table>

**Example**: 4b-MI2(3), L5-R5, C5-NC1; function words removed
Looking into the future
#LancsBox and VR

- Understanding the fabric of language.
- Experiencing language through corpora.
- Pedagogical applications.
Collocations in LLR

The state of play in language learning research

Use: Interest in frequency-based collocations (as part of formulaic language) on the rise; used to assess formulaic L2 production and compare it to L1 users

Method: Identifying collocations in the L2 production; deriving the AM values from a reference corpus (e.g. BNC); adding the values to the L2 production (e.g. Durrant & Schmitt, 2009) and compare to L1 use

Range of AMs: limited. Despite the existence of dozens of AMs - so far only a limited set used in LLR; t-score and MI-score dominant

Rationale for selection: AMs - not fully understood mathematical & linguistics procedures

"it is not clear which of these [MI-score and t-score] (or other) measures is the best to use in research, and to date, the selection of one or another seems to be somewhat arbitrary" (González Fernández & Schmitt, 2015, p. 96)
## The effect of register

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Representativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>British National Corpus (BNC)</td>
<td>98,560,118</td>
<td>Written and spoken (10M), diff. registers</td>
</tr>
<tr>
<td>BNC_Academic</td>
<td>15,778,043</td>
<td>Written, academic writing</td>
</tr>
<tr>
<td>BNC_News</td>
<td>9,412,245</td>
<td>Written, news</td>
</tr>
<tr>
<td>BNC_Fiction</td>
<td>16,143,913</td>
<td>Written, fiction</td>
</tr>
<tr>
<td>BNC – Context governed</td>
<td>6,196,134</td>
<td>Spoken, formal</td>
</tr>
<tr>
<td>BNC – Demographic</td>
<td>4,234,093</td>
<td>Spoken, informal</td>
</tr>
</tbody>
</table>
The effect of register (cont.)

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>Academic</th>
<th>News</th>
<th>Fiction</th>
<th>Formal speech</th>
<th>Informal speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sure</td>
<td>6.8</td>
<td>7.09</td>
<td>7.26</td>
<td>5.78</td>
<td>6.9</td>
<td>6.64</td>
</tr>
<tr>
<td>decision</td>
<td>4.55</td>
<td>3.67</td>
<td>4.07</td>
<td>5.86</td>
<td>6.12</td>
<td>7.91</td>
</tr>
<tr>
<td>point</td>
<td>3.44</td>
<td>2.92</td>
<td>3.84</td>
<td>3.68</td>
<td>4.11</td>
<td>3.12</td>
</tr>
<tr>
<td>Corpus</td>
<td>Size</td>
<td>Representativeness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------</td>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC – Demographic (BNC_D)</td>
<td>4,234,093</td>
<td>Spoken, informal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC – 2014 Spoken (BNC_SP)</td>
<td>4,789,185</td>
<td>Spoken, informal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CANCODE (CANC)</td>
<td>5,076,313</td>
<td>Spoken, informal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Replication?

<table>
<thead>
<tr>
<th></th>
<th>BNC_D</th>
<th>BNC_SP</th>
<th>CANC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>human</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beings</td>
<td>16.3</td>
<td>14.6</td>
<td>14.3</td>
</tr>
<tr>
<td>rights</td>
<td>12.2</td>
<td>11.6</td>
<td>9.4</td>
</tr>
<tr>
<td>nature</td>
<td>10.9</td>
<td>10.7</td>
<td>9.1</td>
</tr>
<tr>
<td><strong>important</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vitally</td>
<td>14.36</td>
<td>13.62</td>
<td>11.28</td>
</tr>
<tr>
<td>terribly</td>
<td>8.39</td>
<td>-</td>
<td>7.28</td>
</tr>
<tr>
<td>very</td>
<td>6.22</td>
<td>5.33</td>
<td>6.03</td>
</tr>
<tr>
<td>really</td>
<td>2.79</td>
<td>3.86</td>
<td>3.54</td>
</tr>
</tbody>
</table>
To address the challenges in LLR

1. **Understand the AMs**: provide rationale for choice of measure, showing understanding of measure, why selected (beyond the fact that it was used by someone before)

2. **Consider a range of AMs and select an appropriate one** to reflect and capture the psycholinguistic concept that you hope to measure & suited to the specific RQ

3. **Consider the effect of genre and topic** (corpus representativeness) in interpretation of the L1 data
thank you much god for goodness much you
References


