The collocation repertoire of EAP users
Ana Frankenberg-Garcia
a.frankenberg-garcia@surrey.ac.uk

Overlapping concepts

prefabricated units
formulaic language
fixed expressions
chunks
collocations
multiword units
idioms
set phrases
lexical bundles

Collocation defined

- Firthian (1957) sense
  - ‘lexical items occurring […] with a greater frequency than the law of averages would lead you to expect’ (Krishnamurthy 1987:70)
- Can be empirically verified against corpus data

Collocations and corpora

Corpus data can provide:
1. A benchmark for what collocations are acceptable in a language or variety of language
   - Dictionaries
     - e.g. Cambridge, COBUILD, Longman, Macmillan, Oxford
   - Textbooks
     - e.g. English Collocations in Use (O’Dell & McCarthy 2008)
   - Tools and resources
     - e.g. Just the Word, Academic Collocations List (Ackermann & Chen 2013)
2. A snapshot of learners’ difficulties regarding collocations (learner corpora)
   - Cambridge Learner Corpus TaLC workshop (Curry & Goodman 2018)

Collocations in learner corpora

- Valuable insights
  - L1 & L2 incongruency problems
  - Preference for high frequency over more unique collocations
    - e.g. VERY cold & BITTERLY cold
- But only a partial picture
  - The fact that learners don’t use a particular collocation doesn’t mean they don’t know it (Gilquin 2007)
  - Short texts & restricted topics, not enough data
    - Studies focus on high frequency words:
      - be, make, top nouns, essay topics
    - In many studies, conclusions come from corpus frequencies alone
    - Dispersion matters: uneven distributions skew the results (Brezina 2018)

Eliciting collocations from writers

- Gap-filling and/or translation tasks
  - To many people, cold food is better than _________ food
    - Farghal and Obiedat (1995:330)
- But poorly designed tests may not elicit the intended data (Schmitt 2010)
- And words participants need to provide not necessarily the words they would want or need to use
- To my knowledge, little research on range of possible collocations writers recall at the moment they are writing
  - VERB + control
    - have control, take control, gain control, seize control, exercise control...
- Limited collocation repertoire can
  - restrict how well you can express yourself
  - disrupt writing process
Research aims

- Gain an understanding of the collocation repertoire available to EAP users
  - Failing to remember academic collocations can disrupt writing
  - Academic writing is cognitively demanding
- Identify patterns in the performance of different types of EAP users
  - L1-English X Other-L1
  - More academic experience X less academic experience
- Focus on
  - number of academic collocations available
  - collocation choices made

Participants

- 90 staff & students from University of Surrey School of Literature and Languages
  - Opportunistic cohort
  - Could be more linguistically aware than average EAP user

<table>
<thead>
<tr>
<th>Role</th>
<th>L1-English</th>
<th>Other-L1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic</td>
<td>8</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>EAP tutor</td>
<td>17</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>PhD student</td>
<td>2</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>MA student</td>
<td>8</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>UG student</td>
<td>21</td>
<td>6</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>34</td>
<td>90</td>
</tr>
</tbody>
</table>

- Sample not balanced

Materials and procedure

- 10 collocation nodes selected from top 50 nouns in Academic Vocabulary List (Gardner and Davies 2014)
  - system, research, effect, data, table, information, change approach, role, analysis, factor
  - Participants likely to know and use them
  - Nouns that evoked a good range of interdisciplinary academic collocations
- Presented in the context of 10 gapped academic writing frames
  - Unlike many gap-filling tests, frames relevant to EAP users
- For each gap
  - Participants asked to supply as many collocates as they could
  - without having to stop and think
  - without going back and revising

Elicitation frames

1. Current research has _______ that…
2. The objective is to _____________ a system that…
3. The data ____________ during the process…
4. The information ____________ in Table 3…
5. They attempted to _______________ the effect of….
6. Another ___________ change observed was…
7. These decisions play an____________ role in….
8. The analysis was___________________ in two stages...
9. An additional factor that _________ these results was…
10. Johns (1991) _________ a different approach to….

Data classification

Focus on lexis

- Different forms of same lemma by the same participant counted only once
  - An additional factor that affected/affects these results
- Spelling mistakes ignored

Data classification cont.

- Words in gaps checked against an expert academic English corpus
  - 37 M word Pearson International Corpus of Academic English (PICAE) (Ackermann et al. 2010)
  - Sketch Engine (Kilgarriff et al. 2014)
- Criteria
  - log Dice ≥ 3
  - ≥ 5 analogous co-occurrences in at least 5 different texts
- Captured very frequent collocations: develop a system
- Sensitive to less common but idiomatic collocations: devise a system
- Left out combinations of words
  - not appropriate in EAP: come up with a system
  - open-choice: discover a system
  - dodgy: hypothesize a system
Overall Results

<table>
<thead>
<tr>
<th>Responses</th>
<th>Lexical Items</th>
<th>Collocations</th>
<th>Not Collocations</th>
<th>Blanks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>23.60</td>
<td>16.44</td>
<td>9.66</td>
<td>3.50</td>
</tr>
<tr>
<td>Median</td>
<td>24.50</td>
<td>18.18</td>
<td>7.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean</td>
<td>25.90</td>
<td>18.27</td>
<td>7.62</td>
<td>0.40</td>
</tr>
<tr>
<td>SD</td>
<td>10.50</td>
<td>7.90</td>
<td>5.00</td>
<td>0.60</td>
</tr>
<tr>
<td>High</td>
<td>70.00</td>
<td>44.00</td>
<td>16.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Low</td>
<td>7.00</td>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- On average 2.59 lexical items per gap (mean values)
- Of which:
  - 1.83 collocations
  - 0.76 not collocations
  - 0.04 blanks

Quantitative findings: L1

<table>
<thead>
<tr>
<th>L1-English</th>
<th>Other-L1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>19.00</td>
</tr>
<tr>
<td>Mean</td>
<td>18.89</td>
</tr>
<tr>
<td>SD</td>
<td>7.89</td>
</tr>
<tr>
<td>High</td>
<td>44.00</td>
</tr>
<tr>
<td>Low</td>
<td>4.00</td>
</tr>
</tbody>
</table>

L1-English & Other-L1 differences were not significant (Mann Whitney, U=861, p<0.05)

Quantitative findings: academic experience

<table>
<thead>
<tr>
<th>Academic Experience</th>
<th>Academics PhD Students MA Students UG Students EAP Tutors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>24.50 20 17 14 09</td>
</tr>
<tr>
<td>Mean</td>
<td>25.36 23.99 16.39 13.33 19.01</td>
</tr>
<tr>
<td>SD</td>
<td>6.80 10.50 4.30 6.50 0.00</td>
</tr>
<tr>
<td>High</td>
<td>41.00 44.00 23.00 28.00 29.00</td>
</tr>
<tr>
<td>Low</td>
<td>12.00 12.00 8.00 2.00 5.00</td>
</tr>
</tbody>
</table>

- Academic experience differences were significant:
  - One-way ANOVA F[3,66] = 11.97, p < .0001
  - Large effect size: $\eta^2 = 0.65$
  - Post-hoc Tukey:
    - academics sig. outperformed undergraduates and MA students
    - PhD students sig. outperformed undergraduates

Caveat...

- Groups not balanced:
  - UG students mostly L1-English
  - PhD students mostly Other-L1
- But for reasonably balanced Academics and MA Students groups:
  - L1-English & Other-L1 differences still not significant
- It cannot be inferred that the groups of higher academic experience performed better because there were more L1-English participants among them
- If anything, quite the opposite was true!

Qualitative analysis

- Only lexical items favoured by at least 20% of each group were taken into account
  - Too much idiosyncratic variation below this threshold
- Different forms of same lemma and spelling variations by different participants grouped together:
  - An additional factor that affected/affects these results
  - The data analysed/analysed during the process

Qualitative findings: L1

1. Current research has shown [proven | demonstrated | found | suggested] that...
2. The objective is to create [develop] a system that...
3. The data [collected | gathered | analysed] during the process...
4. The information [given | provided | presented] in table 3...
5. They attempted to [analyse] the effect of...
6. Another important [significant change observed was...]
7. These decisions play a/an [important | significant | vital | crucial | key] role in...
8. The analysis was done [conducted] carried out in two stages...
9. An additional factor that affected [influences] these results was...
10. Johns (1991) suggests [advocates | proposes] a different approach to...
Qualitative findings: academic experience

1. Current research has shown | proven | suggested | demonstrated | indicated | established that...
2. The objective is to create | develop | design a system that...
3. The data collected | gathered | analysed during the process...
4. The information shown | given | presented | provided | contained in Table 3...
5. They attempted to show | demonstrate | analyse the effect of...
6. Another significant | important | interesting change observed was...
7. These decisions play an | important | crucial | key | significant | vital role in...
8. The analysis was done | carried out | conducted | developed in two stages...
9. An additional factor that affected | influenced | impacted on | contributed to these results was...
10. Johns (1991) suggests | takes | developed | proposes a different approach to...

Discussion & Conclusion

Findings are consistent with

- Hoey’s (2005) Lexical Priming theory
  - People make automatic connections between words once they have encountered them together sufficiently often
  - More exposure to academic register; better recall
- There are no native speakers of Academic English (Hyland 2006; Kossem 2010; Hyland and Shaw 2016; Frankenberger-Garcia et al. 2019)
  - L1-English does not automatically give an advantage to productive collocation repertoire of EAP users
- Hulstijn’s (2011) theory of Higher Language Cognition
  - L1 performance should not be indiscriminately used as a benchmark for assessing L2 proficiency in EAP

Implications for teaching

- Difficulties remembering core academic collocations give credibility to the pedagogical value of general EAP vocabulary tools and resources, e.g.:
  - Academic Word List (Coxhead 2000)
  - Academic Vocabulary List (Gardner and Davies 2014)
  - Academic Collocations List (Ackermann and Chen 2013)
  - Academic Vocabulary in Use (McCarthy & O’Dell 2008)
  - Oxford Learner’s Dictionary of Academic English (Lee et al. 2014)
  - Louvain EAP Dictionary (LEAD) (Granger 2018)
- EAP teaching and reference materials should cater for L1-English students too
  - More in my Friday morning paper:
    - “Developing a tool to help EAP writers with collocations in real time” 😊

Discussion & Conclusion

But

- L1-L2 differences should not go unacknowledged
  - Follow-up analysis of 666 non-collocations elicited in this study could disclose further insights about the effect of L1.
- Core and discipline-specific collocations also need to be considered
  - Are they harder to learn than general EAP collocations?
    - Remember learners get targeted and concentrated exposure to discipline-specific collocations
  - General EAP collocations could actually be harder to notice

Discussion & Conclusion

Main findings

- No major differences in academic collocations L1-English and Other-L1 EAP users remembered
  - Except L1-English users more likely to employ less formal, general English
  - More exposure to non-academic uses of English
- Main variable affecting EAP collocations used was academic experience
  - Academics remembered more collocations
  - Academics agreed more on the same collocations

Discussion & Conclusion

Future research should also explore recall of

- Congruent and incongruent L1/L2 collocations
- Discipline-specific and core academic collocations

Exploratory findings. Should be replicated with

- More participants
- More balanced cohort
- Different disciplinary areas
- Computer-delivered test with screen recording
Thanks to colleagues and students at the University of Surrey School of Literature and Languages who agreed to take part in and/or helped me to collect data for this study.

Data analysis was carried out as part of the ColloCaid project, supported by the Arts and Humanities Research Council (grant number AH/P003508/1).

The collocation repertoire of EAP users