Assisting Writers with Academic English Collocations: lexical data, text editor integration and user feedback

Ana Frankenberg-Garcia
a.frankenberg-garcia@surrey.ac.uk

1. The contradiction was not __________ apparent.
2. They had to __________ the activities of each group.
3. The system has been __________ adopted in the UK.
4. The results are __________ in Table 3.
5. There was a __________ increase ___ prices.
6. Based ___ previous research, we conclude that...

Fill in the gaps

1. The contradiction was not ________ apparent.
2. They had to ________ the activities of each group.
3. The system has been ________ adopted in the UK.
4. The results are ________ in Table 3.
5. There was a ________ increase ___ prices.
6. Based ___ previous research, we conclude that...

About the exercise

1. The contradiction was not __________ apparent.
2. They had to __________ the activities of each group.
3. The system has been __________ adopted in the UK.
4. The results are __________ in Table 3.
5. There was a __________ increase ___ prices.
6. Based ___ previous research, we conclude that...

• What kind of English are the gapped sentences about?
• Can more than one word fit each gap?

Possible answers

1. The contradiction was not clearly, especially, immediately, readily, apparent.
2. They had to monitor, coordinate, organize, regulate the activities of each group.
3. The system has been widely, formally, readily, increasingly adopted in the UK.
4. The results are presented, shown, displayed, provided in Table 3.
5. There was a sharp, slight, significant, steady increase in prices.
6. Based on previous research, we conclude that...

Gaps for testing recall of academic English collocations

Collocation defined

• Firthian (1957) sense
  lexical items occurring together more frequently than just by chance
• How can you check this?
  • Use a corpus

light increase or slight increase

<table>
<thead>
<tr>
<th>Corpus of Academic English</th>
<th>light</th>
<th>slight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence</td>
<td>14,926</td>
<td>1,254</td>
</tr>
<tr>
<td>Co-occurrence 1 left of increase</td>
<td>1</td>
<td>97</td>
</tr>
<tr>
<td>LogDice score</td>
<td>0.7</td>
<td>8.92</td>
</tr>
</tbody>
</table>

Collocations and language learning

• Collocations are notably difficult for L2 learners
  (Pawley and Syder 1983; Natinger and DeCarrico 1992; Howarth 1998; Wray 2002; Frankenberg-Garcia 2018)

• L2 learners often lack awareness of collocation constraints

• Incongruent L1-L2 collocations can be particularly problematic
  (Nesselhauf 2005; Gilquin 2007; Peters 2016)
Which of the following is easier to read?
1. Fine use of collocations can highly improve readability.
2. Effective use of collocations can greatly improve readability.

- Known word combinations are processed with less effort (Conklin and Schmitt 2007, 2012; Ellis et al. 2008)
- Collocationally rich texts are perceived as more fluent (Pawley & Syder 1983; Crossley et al. 2015)

Collocations and academic writing
- L2 users of academic English working outside English-speaking universities show signs of:
  - Limited collocation repertoire
  - Misconceptions about collocation strength
  - Collocation avoidance (Tavares Pinto et al. forthcoming)
- But L1 English undergraduates in the UK also experience collocation difficulties (Frankenberg-Garcia 2018)
- There are no native speakers of academic English (Hyland 2006; Kosem 2010)
- Both undergraduate students and researchers with less experience of writing in English can benefit from academic English collocation assistance

Excellent collocation resources available, but some limitations

EAP textbooks
- Limited information on collocation
- Not practical when writing

General & EAP dictionaries
- Collocation often secondary, hard to find
- Limited coverage of collocation

Collocation dictionaries
- Focus on general English
  - Longman Collocations Dictionary and Thesaurus: appendix with 2469 academic English collocations

Corpora
- Generally not very intuitive
- Noisy data
- Learners
  - Don’t always know how to select appropriate corpus
  - Don’t always know how to build corpus queries
  - Can misinterpret corpus data
  - Can get distracted by corpus data

Corpus-based tools (e.g. Just the Word, SkELL, Flax)
- Not always the best corpus
- Noisy data

What’s more...
- Academic writing is a complex task!
- Looking up collocations can interrupt thoughts and flow of words (Yoon 2016; Tarp et al. 2017)
- Writers will not look things up if are not aware of their limitations
  - Learners tend to overestimate their knowledge of collocations (Frankenberg-Garcia 2005; Laufer 2011)

An alternative solution
What’s different about ColloCaid?
• Can help writers to expand their collocation repertoire
  • Reminders & suggestions of collocations that writers …
    • Don’t remember
    • Don’t feel confident enough to use
    • Are unreasonably avoiding
• Lexicographic data is curated, so you don’t get distracted by
  • Information that is hard to find
  • Irrelevant, misleading or too much information

What’s different about ColloCaid?
• Based on data-driven learning (Johns 1991)
  • Collocations are shown, not explained
  • Learners value examples (Chan 2011; Dziemianko 2006, 2012)
  • Multiple examples help more (Frankenberg-Garcia 2012; 2014; 2015)
• Interactive text-editor integration
  • Works in real time, writers do not have to stop writing
  • Informed by HCI research (Rees et al. 2019)

Methodology: where to start?
( ) Where in my text can I fit in conduct research?
(x) What verb can I use with research?
• Starting point is the collocation node, not the collocation
• 500 maximally useful general academic English nodes
  • 36-month project
  • Nouns, verbs and adjectives (not adverbs)
• Nodes selected from 3 well-established word lists of general academic English vocabulary
  • Extracted using different methods and different corpora
• Build on the strengths of each list

Methodology: academic word lists
1. AKL: Academic Keyword List (Paquot 2010)
  • academic keyness
2. AVL-BAWE: subset of Academic Vocabulary List (Gardner & Davies 2014) that overlaps with BAWE student writing word list (Durrant 2016)
  • novice academic writing vocabulary
3. ACL: Academic Collocations List (Ackermann & Chen 2013), using as nodes the 704 headwords listed in the appendix to Longman Collocations Dictionary (Mayor 2013)
  e.g. for conduct research, headword/node = research
  • strong collocations

Methodology: nodes in at least 2 lists
AVL-BAWE 387
ACL 704

514

Methodology: which collocations?
Collocates likely to be looked up for each node:
• Lexical collocations (not discipline-specific)
  • N-V system: design, develop, create, establish...
  • N-Adj potential: full, huge, great, immense, future....
  • V-Adv affect: directly, significantly, negatively, positively...
  • Adj-Adv acceptable: mutually, morally, ethically, perfectly...
  • Grammatical collocations
  • V-prep approach: to (study, analysis, problem)
  for (investigating, identifying)
  towards (discipline, goals)
Methodology: which sources?

Expert academic English corpora
• BNC academic
• COCA academic
• Pearson International Corpus of Academic English
• Oxford Corpus of Academic English
  • 70 million words
  • Published academic papers and books by academic expert writers
  • Not necessarily L1 English
  • 26 disciplines
  • From biochemistry to sociology

Methodology: which tools?

- Sketch Engine (Kilgarriff et al. 2004, 2014)
- Word Sketches
  • Collocates sorted per grammar relation
    • e.g. ADJ + research
  • Co-occurrence
  • LogDice score

Methodology: which collocates?

- LogDice score: ≥ 5, ranked
- Co-occurrence: ≥ 10 lexical; ≥ 100 grammatical
- Dispersion: ≥ 3 disciplines

Methodology: which examples?

- From expert writing corpora
  • Light editorial intervention
  • Short
  • Less effort to process
  • Save screen space
  • Transferable to other disciplines
  • Promote data-driven learning
- Top eight
  • From expert writing corpora
  • Light editorial intervention
  • Short
  • Less effort to process
  • Save screen space
  • Transferable to other disciplines
  • Promote data-driven learning
- GDEX (Kilgarriff et al. 2008)

Methodology: prototype development

- Writers have different preferences
  • Windows + MS Word
  • Mac + LaTeX
  • Windows + TinyMCE
    • Widely used, open-source editor
    • Use on browser
    • Compatible with multiple devices and operating systems
    • Familiar interface
    • Great for prototype testing

Version 0.6 lexicographic database

- 561 academic lemmas
- 32,186 collocation suggestions
- 28,581 curated corpus examples

(all collocations also linked to non-curated examples from SkELL)
方法论：人机交互

Fitts's law (Fitts 1954)
时间去获得一个目标是其长度和大小的函数
- 分词用户通常需要覆盖一个长距离来找到一个共现
  - 停止写作
  - 去词典
  - 查看词头
  - 转到共现部分
  - 识别目标共现
  - 回到写作
- 语料库用户同样
  - 停止写作
  - 去语料库
  - 选择查询类型
  - 执行查询
  - 解释原始语料库数据
  - 回到写作

方法论：人机交互

工作记忆
“平均人可以记住7±2个东西” (Miller, 1956)
- 在ColloCaid中，交互菜单最多不超过八个项目

方法论：人机交互

心理模型 (Nielsen 2010)
用户对一个系统的想法严重地影响了他们如何使用它
- 心理模型的混淆
  - 设计者和用户的心理模型之间的差距
  - 在浏览器中
    - 错误地将地址栏视为搜索引擎
  - 在字典中
    - 错位的地方找词语
    - e.g. A stitch in time saves nine

方法论：人机交互

心理模型的混淆
- 在早期的ColloCaid原型中，一些用户认为被划线的词是错误的
- 在后期的版本中，黑色下划线变为了绿色

系统易用性量表 (Brooke, 2013)
一个快速和粗放的工具来测量系统的易用性（硬件、软件、网站等）
- 十个交替的正反陈述关于系统的易用性
  - 用户在利克特量表上评分
  - 转换为0-100的分数
  - 标准的易用性测试
  - 被1300多篇引用
  - 快速和便宜
  - 可靠，即使样本量小
  - 合法 - 它测量它声称测量的东西
**Methodology: Human-Computer Interaction**

**System Usability Scale**

1. How strongly do you agree or disagree with the usability statements below:

   - I think that I would like to use ColloCaid frequently.
   - I found ColloCaid unnecessarily complex.
   - I think that ColloCaid suits my way of use.
   - I have received too much support from support personnel to be able to use ColloCaid.
   - The various functions in ColloCaid were well integrated.
   - I thought with too much inconsistency in ColloCaid.
   - I would imagine that most people would learn to use ColloCaid very quickly.

**Initial user studies**

- Version 0.1 Poland & France
- Version 0.2 Brazil
- Version 0.3 Brazil & Spain
- 122 participants
  - L1 Polish, French, Portuguese & Spanish
  - MA & PhD students, Research Fellows & Academics
    - Linguistics, Computer Science, Mathematics, History, Sociology, Astronomy, Biology, Physics...
- System Usability Scale (SUS) questionnaire
- Plus specific questions about ColloCaid

**SUS results for versions 0.1 to 0.3**

Over 20k collocation suggestions & links to external examples added since then

**Suggestions for improvement**

More words could be added to ColloCaid, like academic words from specific areas
- Database has been expanded since
- Deliberately not covering subject-specific terms

Maybe it could be compatible with text editors we use daily, such as Microsoft Word
- Couldn’t agree more!

Install an auto saving mechanism so that the text that I am composing is not damaged or lost
- Privacy & data protection issues
- Export to txt function to be added

The interface / appearance could be more appealing but that is not a really important issue and I guess it is the last thing to improve
- We’re on it!

**Some positive feedback**

- It is user friendly and it has lots of examples
- It’s very intuitive. Easy to use. I like it when it gives only 1 example to begin with and then later you can get more, if you need more
- You can look for the word you’re interested in without wasting time
- I like the fact that it gives me combinations of words that sound more formal and academic than the ones I would think myself

**Visualisation: tree view**
Who is using ColloCaid?
- Prototype first released to the public at eLex 2019
  - Not broadly advertised
- Current version 0.5
- Visualization options & export function soon to come in version 0.6
- Still some bugs

1085 registered users on 18/05/2020

How did they find out about ColloCaid?

Intended uses

Main activity of users

PHD Student
Master Student
Undergraduate Student
University lecturer/professor
Other
Research fellow
Teacher of academic English
Secondary School Student
ESL Teacher
Secondary School Teacher
Main writing language of users

Slovak Afikannis Albanian Greek
Afar Arabic Armenian Azerbaijani Bengali
Bulgarian Catalan Chinese
Croatian Danish Dutch
Estonian Farsi Finnish
French Gaelic German
Greek Hungarian Italian
Japanese Korean
Latin Lithuanian
Latvian Norwegian
Persian Polish
Portuguese Romanian
Russian Serbian
Slovak Slovenian
Spanish Swahili
Swedish Tajik
Turkish Urdu
Uzbek
Welsh
Yiddish

Next steps...

- Address known bugs
- Alternative visualization trials
- More usability tests

Try our prototype

www.collocaid.uk
ColloCaid is funded by the
UK Arts and Humanities Research Council
(AH/P003508/1)

Principal Investigator
Dr Ana Frankenberg Garcia (Surrey)
Co-Investigators
Prof Robert Lew (Poznan), Prof Jonathan Roberts (Bangor)
Researchers
Dr Geraint Rees (Surrey), Peter Butcher (Bangor) & Dr Nirwan Sharma (formerly Bangor)