ColloCaid: A real-time tool to help academic writers with English collocations

Robert Lew, Ana Frankenberg-Garcia, Geraint Rees, Jonathan Roberts, and Nirwan Sharma
EURALEX 2018
Away from static dictionaries

- the *digital revolution*
- traditional stand-alone dictionaries are on their way out
- dictionaries as we know it giving way to digital tools
- geared towards specific tasks
Collocation (for Academic English)

- Appropriate collocations make texts more idiomatic, more readable (Hoey 2005, Ellis et al. 2008; Conklin and Schmitt 2012)

- Collocation has always been a particular area of difficulty for language learners (Pawley and Syder 1983; Natinger and DeCarrico 1992; Howarth 1998; Wray 2002)

- Language users often lack awareness of collocational constraints (Nesselhauf 2005; Durrant and Schmitt 2009; Laufer and Waldman 2011)

- Importance of English as a lingua franca for the dissemination of science (Jenkins 2014)

- There are no native speakers of academic language (Hyland 2006, Kosem 2010, Frankenberg-Garcia 2017)
EXISTING RESOURCES FOR COLLOCATION: DICTIONARIES
LDOCE Collocation box

COLLOCATIONS

VERBS

do/carry out research (also conduct research formal)
  The research was carried out by a team of scientists at Edinburgh University.
  Little research has been conducted into the subject.

undertake research formal (=start or do research)
  They are planning to undertake research into the genetic causes of the disease.

ADJECTIVES

scientific research
  Our conclusions are based on scientific research.

medical research
  The charity raises money for medical research.

historical research
  This is a fascinating piece of historical research.

basic research (=the most important or most necessary area of research)
  He wants to conduct basic research into the nature of human cells.

extensive research (=research that examines a lot of information and details)
  Extensive research has proved the theory wrong.

painstaking research (=very careful and thorough research)
  She spent years carrying out painstaking historical research.

pioneering research (=research that produces completely new information)
  Watson did pioneering research on the long-term effects of dieting.
Collocation Dictionaries
Macmillan Collocations Dictionary

- using particular methods academic, applied, comparative, empirical, qualitative, quantitative
  Much of the research is qualitative and based on both male and female samples.
- v+N do research carry out, conduct, do, perform, pursue, undertake She is continuing to do research into medieval European dress.
- provide money for research finance, fund, sponsor, support Every penny received goes directly to buy equipment or to fund research.
- officially ask someone to do research commission The government has commissioned further research to be carried out by The National Parenting and Families Institute.
- tell people the results of research deliver, present, publish, report Professor David Mellor is to present his research at a conference organised by Compassion in World Farming next month.
- be in charge of a programme of research coordinate, direct, lead, manage, supervise The research was led by Professor Dennis Parker.
- look at other people’s research assess, evaluate, review This paper will review the research of the past decade and consider some of the gaps in our knowledge.
- report or publish your research disseminate, present, publicize, publish, report The monthly meeting provides a forum for team members to present their research.
EXISTING RESOURCES: CORPORA
**SkELL**  (Baisa & Suchomel 2014)

![SkELL interface with the word 'research' entered](image)

- **Research** as noun
  - Verbs with research as subject: suggest, show, focus, indicate, reveal, demonstrate, find, involve, examine, aim, support, confirm, investigate, identify, prove
  - Verbs with research as object: conduct, fund, undertake, do, publish, focus, support, need, perform, pursue, sponsor, pioneer, promote, apply, continue
  - Adjectives with research: ongoing, assistant, underway, online, relevant, necessary, vital, essential, important, critical, first, available
research + suggest  2.09 hits per million

1. Some research suggests moods are related to overall job satisfaction.
2. Additional research suggests the social component of ritual.
3. Further research suggests cosmic dust might have skewed the results.
4. Other research suggests that later training is associated with problems.
5. Some research has suggested that exercise may have antidepressant effects.
6. This research suggests independence of episodic and semantic musical memory.
7. Some research suggests alternate estimates to the accepted workforce size.
8. Some research suggests an association between personality and job satisfaction.
9. The same research suggests they may have fought barefoot.
10. Recent research suggests that this is untrue.
**Learning Collocations**

```
Flax

Interactive language learning

search in academic English (BAWE)

Family words: researched, researcher, researchers, researches, researching

<table>
<thead>
<tr>
<th>used as a noun</th>
<th>research project</th>
<th>303</th>
<th>research contract</th>
<th>103</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>research programme</td>
<td>237</td>
<td>research work</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>market research</td>
<td>209</td>
<td>research students</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>research findings</td>
<td>141</td>
<td>research interests</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>research team</td>
<td>130</td>
<td>research methods</td>
<td>80</td>
</tr>
</tbody>
</table>
```
flax
Limitations

- Corpora directly: require skills most don’t have
- If curated (mostly dictionaries): limited coverage, examples
- Information may be hidden away within the entry
- If not curated (SkELL, flax, netspeak.org, other corpus-based): noise
- Not specific to Academic English (exceptions: LEAD, OLDAE)
- Users have to interrupt their writing to consult the collocation resource
WRITING ASSISTANTS
A misspelled word.

A repeated word.

A simple grammar mistake.

A similar word.
Cambridge English
Write & Improve

It’s simple

1) Choose a topic
2) Write about it in English
3) Get instant feedback
4) Improve!

Write & Improve is a free service for learners of English to practise their written English. Submit your written work and receive feedback in seconds, covering spelling, vocabulary, grammar and general style.

Start writing →

The exciting new way to improve your English
Your writing, at its best.

Millions of students and professionals use Grammarly to check their papers, emails, and other important documents.

Grammarly will make sure
In the present contribution, we explain the patterns

<table>
<thead>
<tr>
<th>less</th>
<th>more patterns</th>
<th>less</th>
<th>more examples</th>
<th>write</th>
<th>edit</th>
<th>English</th>
</tr>
</thead>
</table>

[V] **present something in → present something** 48
assembly [presented the constitution in] 24, 1979, => assembly [presented the constitution] October 24, 1979,
[presented a Norwegian lumberjack in] Upper Midwest =>, [presented a Norwegian lumberjack] the Upper Midwest

[V] **present something of → present something** 28
able to [present proof of] the presumed licensee has => able to [present proof] the presumed licensee has
Doyle [presents a comparison of] debuting character => Doyle [presents a comparison] his debuting character

[V] **present with something → present something** 26
and [presents with a significantly worse deformity] postural kyphosis => and [presents a significantly worse deformity] postural kyphosis
often [present with an internalized, inattentive sub-type] the disease which => often [present an internalized, inattentive sub-type] the disease which
Existing writing assistants: problems

- Mostly *reactive* feedback (not real-time)
- Often corpus-based using NLP: noise
- Not specific to Academic English
- Not about collocation
- Usability issues: intrusive, unintuitive
OUR SOLUTION

Project led by University of Surrey (A. Frankenberg-Garcia, G. Rees)
with Bangor University (J. Roberts, N. Sharma)
Adam Mickiewicz University in Poznań (R. Lew)
Funded by Arts & Humanities Research Council UK (AH/P003508/1)
An alternative solution

- An intuitive collocation-aware lexicographic resource integrated with digital writing environments that:
  - *requires little training or instruction*
  - *raises language awareness*
- Focus on Academic English (EAP) collocations:
  - *carry out + research*
  - *statistically + significant*
  - *further + studies*
Our set of collocate bases (nodes)

- Based on three recent research-based sets of EAP vocabulary
  - Academic Vocabulary List (AVL, Gardner & Davies 2014, based on COCA); narrowed down as per Durrant (2016), who found that just 427 AVL items were frequent in a broad range of disciplines in university student writing (BAWE corpus)
  - Academic Keyword List (AKL, Paquot 2010), based on the LOCNESS Corpus (Granger 1998) of British and American written assignments
  - Academic Collocation List (ACL, Ackermann & Chen 2013) found in the appendix to the Longman Collocations Dictionary

- Criterion for inclusion: item present in more than one of these 3 lists
At least two out of three lists: 512 nodes
512 nodes selected

Noun bases
282
- Adjectives & modifiers:
  - e.g. a complex system, a classification system, an existing system, etc.
- Verbs (object of):
  - e.g., design a system, develop a system, establish a system, etc.
- Verbs (subject of):
  - e.g. the system consists of, the system functions, the system provides, etc.
- Prepositions:
  - e.g. within a system, a system for something

Verb bases
136
- Adverbs:
  - e.g. undoubtedly contribute, contribute significantly, contribute substantially, etc.
- Prepositions:
  - e.g. contribute to

Adjective bases
94
- Adverbs:
  - e.g. statistically significant, particularly significant, etc.
important (adjective)
Oxford Corpus of Academic English (April 2012) freq = 55,647 (0.69 per million)

<table>
<thead>
<tr>
<th>ADV</th>
<th>ADJ*</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>particularly +</td>
<td>917</td>
<td>10.94</td>
</tr>
<tr>
<td>equally +</td>
<td>401</td>
<td>10.40</td>
</tr>
<tr>
<td>especially +</td>
<td>430</td>
<td>10.35</td>
</tr>
<tr>
<td>increasingly +</td>
<td>398</td>
<td>10.17</td>
</tr>
<tr>
<td>very +</td>
<td>1,495</td>
<td>10.03</td>
</tr>
<tr>
<td>extremely +</td>
<td>796</td>
<td>9.75</td>
</tr>
<tr>
<td>as +</td>
<td>839</td>
<td>9.52</td>
</tr>
<tr>
<td>so +</td>
<td>361</td>
<td>9.31</td>
</tr>
<tr>
<td>critically +</td>
<td>124</td>
<td>9.13</td>
</tr>
<tr>
<td>vitally +</td>
<td>118</td>
<td>9.11</td>
</tr>
<tr>
<td>potentially +</td>
<td>138</td>
<td>8.79</td>
</tr>
<tr>
<td>crucially</td>
<td>78</td>
<td>8.52</td>
</tr>
<tr>
<td>really</td>
<td>75</td>
<td>8.19</td>
</tr>
</tbody>
</table>
Adding collocates and examples

- logDice score of ≥ 5
- minimum co-occurrence frequencies
  - 10 for lexical collocates
  - 100 for prepositions
- remove collocates that are too general (e.g. own and good)
- remove collocates restricted to a small number of disciplines (e.g. entrepreneurial ability restricted to business studies)
- select three examples per each collocate-base pair
- replace proper nouns with pronouns, change date, numerical and name references
- editing done in a cloud-based spreadsheet
<table>
<thead>
<tr>
<th>BASE</th>
<th>POS</th>
<th>RELATION</th>
<th>COLLOCATE</th>
<th>CO-</th>
<th>ASSOCI</th>
<th>EXAMPLE1</th>
</tr>
</thead>
<tbody>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>roughly</td>
<td>108</td>
<td>11.42</td>
<td>the latter two groups had roughly equal rates of break</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>exactly</td>
<td>62</td>
<td>10.85</td>
<td>total costs and our total revenues are exactly equal</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>nearly</td>
<td>57</td>
<td>9.89</td>
<td>three experiments were performed using nearly equal</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>almost</td>
<td>69</td>
<td>9.26</td>
<td>men and women are apparently almost equal now in t</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>formally</td>
<td>10</td>
<td>8.35</td>
<td>this occurs where treatment is formally equal</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>necessarily</td>
<td>14</td>
<td>7.81</td>
<td>attachment does not necessarily equal ownership</td>
</tr>
<tr>
<td>equal</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>relatively</td>
<td>20</td>
<td>5.48</td>
<td>all household members have relatively equal access to c</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>particularly</td>
<td>912</td>
<td>10.94</td>
<td>the sensitivity of a test is particularly important</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>equally</td>
<td>401</td>
<td>10.4</td>
<td>the two stages are equally important and interlinke</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>especially</td>
<td>430</td>
<td>10.35</td>
<td>especially important were the localization of brain and</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>increasingly</td>
<td>398</td>
<td>10.17</td>
<td>public relations is becoming an increasingly important</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>very</td>
<td>1495</td>
<td>10.03</td>
<td>it is very important for an economy to be stable</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>extremely</td>
<td>296</td>
<td>9.75</td>
<td>they see work and its consequences as extremely imp</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>critically</td>
<td>124</td>
<td>9.13</td>
<td>determine which points of critically important informa</td>
</tr>
<tr>
<td>important</td>
<td>j</td>
<td>ADV ADJ*</td>
<td>vitally</td>
<td>118</td>
<td>9.11</td>
<td>the link between the two concepts is therefore vitally</td>
</tr>
</tbody>
</table>
Example scenarios

- Will help EAP users with collocations in real time (Frankenberg-Garcia et al. 2019)
- Raising awareness of collocation problems

A small increase of temperature

× Do not say increase of something
√ Say increase in something
Example scenarios

- Not just correcting reactively, but guiding the process

Research done as a result of...
- conduct research
- research suggests
- empirical research
- research on something
- conduct research
- carry out research
- undertake research
- **research carried out** in the 1990s identified...
- before any *research* is *carried out*, it is essential that...
- this may mean that *research* can only be *carried out* in the dry season
Example scenarios

- Customizable
- Option to switch off real-time help and check text at the end
- We plan to run usability tests on a prototype

Research done as a result of...

- conduct research +
- research suggests +
- empirical research +
- research on something +
First prototype implementation

- Custom plugin for the TinyMCE text editor
  - *TinyMCE* is
    - open-source
    - lightweight
    - extensible
    - written in JavaScript
    - intended for web applications
    - client-side (web browser does the work)
PROTOTYPE DEMO

Local File | On Vimeo | Live Demo | Screencast
How will ColloCaid be different?

- Lexicographic data will be curated so that we provide just the right amount (less is more)
- Specific focus on Academic English
- When writing, users won’t get distracted with:
  - irrelevant or misleading information
  - information overload
- Prompts will be discreet and customizable, so that
  - they don’t fill too much screen space
  - users don’t have to deal with lengthy explanations
- Metalanguage will be avoided, so users don’t have to struggle with grammatical terminology
The ColloCaid team
(in the middle of a Skype conference)