COPYEDITING AND CORPUS LINGUISTICS

Jonathon Owen

ACES 2016
What is a corpus?

- A corpus (plural corpora) is a collection of electronic text compiled for research purposes.
- Like other researchers, linguists need data, and digitizing a whole bunch of text is a good way to get some.
- The words are usually tagged by part of speech to make searching easy.
A brief history of corpora

- Corpora were originally created and used primarily by researchers, and early corpora were typically only a few million words at best.

- In the ’60s, the first *American Heritage Dictionary* used a corpus to give it a solid empirical basis.

- In the early 2000s, BYU’s Mark Davies started making publicly available corpora in the 400–500 million word range.

- In 2010 Google published the Ngrams Viewer with 155 billion words.
How are corpora used?

- Linguistics: researching word frequency, concordances and collocations (which words occur together), and variation and change
- Language teaching: seeing how natives actually say it, seeing which words are the most common
- Translation: comparing equivalent constructions in different languages
- Lexicography: seeing how words are used in context, discovering collocates, examining different senses
  - **Without data, a lexicographer is just someone sitting at a keyboard typing all the words they know.**
Why should an editor care about corpus linguistics?

- Because usage dictionaries and style guides aren’t always up to date, and they can’t cover every issue.
- Because even the issues that they do cover might not be accurate if they’re not based on empirical evidence.
- Because sometimes it’s hard to see past our own biases, and sometimes our intuitions are not reliable.
- Because science!
Don’t worry—this isn’t “anything goes”

- Most corpora are based on published materials, which means that the text has generally been edited.
- Like dictionaries, corpora can provide facts, but you’ll still have to exercise your own judgment in the end.
- And anyway, if the fact that everybody does it doesn’t make it right, what does make it right?
What are some popular corpora?

- Corpus of Contemporary American English (COCA)
- Corpus of Historical American English (COHA)
- Google Books Ngrams Viewer
Corpus of Contemporary American English (COCA)

- [http://corpus.byu.edu/coca](http://corpus.byu.edu/coca)
- 520 million words
- 1990–present (text is continually added)
- organized by genre
  - spoken
  - fiction
  - magazines
  - newspapers
  - academic
Corpus of Historical American English (COHA)

- http://corpus.byu.edu/coha
- 450 million words
- 1800–present (text is continually added)
- organized by genre
  - fiction
  - magazines
  - newspapers
  - nonfiction
  - academic
Google Books Ngrams Viewer

- [http://books.google.com/ngrams](http://books.google.com/ngrams)
- 155 billion words
- 1800–present* (text is occasionally added)

*technically 1500s–present, but the 1500–1800 data is mostly garbage
Corpus comparison

- **COCA/COHA**
  - powerful but unintuitive interface
  - data can be copied and pasted into spreadsheet (but not exported directly)
  - text is balanced across years and genres

- **Google Books**
  - super-simple but less-powerful interface, a lot of features are buried
  - data can be viewed but not copied or exported
  - text is not balanced across years and genres
Searching in COCA & COHA: display

- **List**
  - not always the most useful—just lists all the search results with their frequencies
  - can be useful for comparing all the results of searches with wildcards or part-of-speech tags

- **Chart**
  - creates bar graphs showing the frequencies by genre and by year range
  - great for quickly comparing usage in different genres or change across time

- **KWIC (keyword in context)**
  - great for highlighting which words or parts of speech typically follow the search term
  - shows 100 random hits

- **Compare**
  - I’m not gonna lie—I don’t use this one because I can never seem to get it to work
Searching in COCA & COHA: search string

- **Word(s)**
  - the main term or terms you’re searching for (not case sensitive)
  - can be one or more words, including wildcards and part-of-speech tags

- **Collocates**
  - search for words that occur within a certain range of the main search term
  - the two drop-downs search before and after the main search term—by default it looks 4 words before and 4 words after

- **POS List**
  - no, it stands for “part of speech,” not the other thing
  - lets you search by part of speech and some subcategories (different verb forms, plural vs. singular nouns, positive, comparative, and superlative adjs., etc.)
Searching in COCA & COHA: search query syntax

- all inflected forms of a word: put it in brackets—[word]
- a word only as a particular part of speech: put a POS tag after it—word.[v*]
  - mind the period—it’s necessary to apply the tag to that word
- synonyms of a word: [=word]
- or: separate terms with vertical bar—word|term|phrase
- wildcards: * for any number of letters or a whole word, ? for exactly one letter
- not: minus sign followed by search term
Searching in Google Books

- Put your search term or terms in the box
- Specify a date range if you don’t want the default 1800–2000
- Choose a corpus (you’ll probably want English or American English)
- For information on more advanced searches, go [here](http://example.com) (or click “About Ngram Viewer” in the footer).

Graph these comma-separated phrases: Albert Einstein, Sherlock Holmes, Frankenstein

between 1800 and 2000 from the corpus English with smoothing of 3

[Search lots of books]
Searching in Google Books: search query syntax

- Wildcards: *
  - Only searches words, not parts of words, and only lists top 10 results

- Inflected forms: _INF
  - Example: walk_INF = walks, walking, walked

- Part-of-speech tags
  - Can be combined with _INF tag
  - Can stand alone or be appended to a word

**Note:** you cannot mix wildcards and inflection or part-of-speech tags in one search term

- Start or end of sentence: _START_ and _END_

- A word as a modifier: word=>modifier

- Search by a particular corpus: term followed by colon and tag for corpus (eng_us_2009, eng_2012, etc.)
Searching in Google Books: doing math with search queries

- You can add, subtract, multiply, and divide search queries—just use +, -, *, and /, and use parentheses as necessary to group.

- The first example below simply compares two usages; the second looks at the percentage of the time one of the two is used.
  - composed of, comprised of
  - comprised of/(composed of + comprised of)
Researching editing questions

wow

such usage

very language

many grammar
all right/Alright

- Google Books
- COHA
- COCA
data is/data are

- Google Books
- COHA
- COCA
e-mail/email

- Google Books
- COHA
- COCA
impacted

- Google Books
- COCA
- COHA
Internet/internet

- Google Books
less/fewer

- less [plural noun] COHA
- fewer [plural noun] COHA
- less than [number] [plural noun] COHA
- fewer than [number] [plural noun] COHA
- less than [number], fewer than [number] Google Books
- [noun] or less, [noun] or fewer Google Books
regardless/irregardless

- Google Books
- COHA
- COCA
sneaked/snuck

- Google Books
- COCA
that/which

- **that/which** Google Books American
- **that/which** Google Books British
toward/towards

- Google Books
- COCA
who/whom

- Google Books
- COHA
- COCA
Journalese

- Temblor
- Oust/Ouster
- Garner
- Woes
- Lambaste
Some pitfalls of corpus searches

- A corpus search only tells you about the nature of the corpus
- Sometimes the data is skewed or unreliable in some way—you may have to dig deeper to see if it holds up
- Data can’t tell you what you should or should not do
- A couple of examples of misleading results:
  - **e-mail, email**
    - The pre-1990s results are for the unrelated and obsolete word *email*, meaning “enamel.”
  - **the poop spike**
    - People weren’t especially interested in poop in the 1920s—the results are all about boats, and the spike is probably a result of unbalanced data across years and genres.
Results from my master’s thesis

- The two most popular usage changes made by editors:
  - *which* > *that*
  - *towards* > *toward*

- The apparent dominance of *toward* and of *that* as a restrictive relative pronoun are an artifact of copyediting—editors have been hunting *towards* and *which* to extinction.

- Conclusion: It’s really easy to get caught in feedback loops between editing and lexicography—sometimes we drift away from what everyone else is doing.
Conclusion

- While corpora won’t ever replace traditional references, they can supplement them in some really great ways.
- Corpus data can help you combine the best aspects of prescriptivism and descriptivism.
- Good editing is informed editing.
Questions or comments?

- Feel free to contact me
  - On Twitter: @ArrantPedantry
  - On the web: www.arrantpedantry.com
  - By email: jonathon@arrantpedantry.com