Copyediting and Corpus Linguistics

JONATHON OWEN

ACES 2017
• “Through learning and practicing traditional seafaring and wayfinding, they discover and **ingrain** the ethos and values inherent to Pacific Island cultures.”
Ingrain

• *Merriam-Webster’s definition:*

• **Definition of INGRAIN**
  
  *transitive verb*
  
  : to work indelibly into the natural texture or mental or moral constitution

  
  <the journalism professor has long *ingrained* his students with a deep respect for their chosen profession>

  
  <the third-world privation he had witnessed forever *ingrained* itself upon the young doctor's memory>
only learn faster but the memories last longer. Like, they're more deeply ingrained

idea that I would say good morning to someone - and it's just so ingrained in who I am.

Do you think that racism is so ingrained in some corners of our society that it will just never be done away with

No. It is deeply ingrained, and there are subtle forms of discrimination all around us, and we’re

Unshakeable stoicism was ingrained into women of Helena's class.

I remember the procedures for egress, an exit strategy ingrained over so many years of navy training.

Melody struck the keyboard, her muscle memory ingrained with the instinct to perform

That moment was still ingrained in my consciousness. It was my first experience recognizing my own stupidity.

He feels the grit ingrained into the skin of his palms, the dust on his hands that never quite

get people to not see the poor as different from other people. That was ingrained early for me.
Why use a corpus?

• Because dictionaries can’t tell you everything there is to know about how to use a word

• 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.

• Doing research in corpora is a way to determine what Standard English is empirically.

• 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 is a corpus?

- A corpus (plural *corpora*) is a database of electronic text compiled for research purposes.
- Like other researchers, linguists and lexicographers 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.
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
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
• 400 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
Simple frequency searches
all right/alright

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

- Google Books
- COHA
- COCA
data is/data are
e-mail/email

- Google Books
- COHA
- COCA
Internet/internet

- Google Books
regardless/irregardless

- Google Books
- COHA
- COCA
sneaked/snuck

- Google Books
- COCA
toward/towards

- Google Books
- COCA
who/whom

• Google Books
• COHA
• COCA
Journalese

- Temblor
- Oust/Ouster
- Garner
- Woes
- Lambaste
Searching with part-of-speech tags
acquiesce in/acquiesce to

- Google Books
- COCA

- Tip for Google Books: Right-click to combine or split different forms of the same headword.
impact as a verb

- Google Books
- COCA
- COHA
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
that/which

- that/which Google Books American
- that/which Google Books British
Advanced search stuff
Part-of-speech tags in Google Books

<table>
<thead>
<tr>
<th>Part-of-Speech Tag</th>
<th>Tag Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td><em>NOUN</em></td>
</tr>
<tr>
<td>verb</td>
<td><em>VERB</em></td>
</tr>
<tr>
<td>adjective</td>
<td><em>ADJ</em></td>
</tr>
<tr>
<td>adverb</td>
<td><em>ADV</em></td>
</tr>
<tr>
<td>pronoun</td>
<td><em>PRON</em></td>
</tr>
<tr>
<td>determiner or article</td>
<td><em>DET</em></td>
</tr>
<tr>
<td>an adposition: either a preposition or a postposition</td>
<td><em>ADP</em></td>
</tr>
<tr>
<td>numeral</td>
<td><em>NUM</em></td>
</tr>
<tr>
<td>conjunction</td>
<td><em>CONJ</em></td>
</tr>
<tr>
<td>particle</td>
<td><em>PRT</em></td>
</tr>
<tr>
<td>root of the parse tree</td>
<td><em>ROOT</em></td>
</tr>
<tr>
<td>start of a sentence</td>
<td><em>START</em></td>
</tr>
<tr>
<td>end of a sentence</td>
<td><em>END</em></td>
</tr>
</tbody>
</table>

These tags can either stand alone (_PRON_) or can be appended to a word (she_PRON)

These tags must stand alone (e.g., _START_)
# Part-of-speech tags in BYU corpora

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Original</th>
<th>New (word)</th>
<th>New (abbrev)</th>
<th>CQP-like</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common nouns</td>
<td>[nn*]</td>
<td>NOUN</td>
<td>N</td>
<td>_nn</td>
<td>sun, love</td>
</tr>
<tr>
<td>Proper nouns</td>
<td>[np*]</td>
<td>NAME</td>
<td>NP</td>
<td>_np</td>
<td>John, Chicago</td>
</tr>
<tr>
<td>Common and proper nouns</td>
<td>[n*]</td>
<td>NOUN+</td>
<td>N+</td>
<td>_n</td>
<td>sun, Sonny</td>
</tr>
<tr>
<td>Lexical verb (no do, be, have)</td>
<td>[vv*]</td>
<td>VERB</td>
<td>V</td>
<td>_vv</td>
<td>decide, jumped</td>
</tr>
<tr>
<td>All verbs (incl do, be, have)</td>
<td>[v*]</td>
<td>VERB+</td>
<td>V+</td>
<td>_v</td>
<td>decide, has, is</td>
</tr>
<tr>
<td>Adjectives</td>
<td>[j*]</td>
<td>ADJ</td>
<td>J</td>
<td>_j</td>
<td>nice, clean</td>
</tr>
<tr>
<td>Adverbs</td>
<td>[r*]</td>
<td>ADV</td>
<td>R</td>
<td>_r</td>
<td>soon, quickly</td>
</tr>
<tr>
<td>Pronouns</td>
<td>[p*]</td>
<td>PRON</td>
<td>_p</td>
<td>she, everyone</td>
<td></td>
</tr>
<tr>
<td>Prepositions</td>
<td>[i*]</td>
<td>PREP</td>
<td>_i</td>
<td>from, everyone</td>
<td></td>
</tr>
<tr>
<td>Articles</td>
<td>[a*]</td>
<td>ART</td>
<td>_a</td>
<td>the, his</td>
<td></td>
</tr>
<tr>
<td>Determiners</td>
<td>[d*]</td>
<td>DET</td>
<td>_d</td>
<td>these, all</td>
<td></td>
</tr>
<tr>
<td>Conjunctions</td>
<td>[c*]</td>
<td>CONJ</td>
<td>_c</td>
<td>that, and, or</td>
<td></td>
</tr>
<tr>
<td>Negation</td>
<td>[x*]</td>
<td>NEG</td>
<td>_x</td>
<td>not, n’t</td>
<td></td>
</tr>
<tr>
<td>Numbers</td>
<td>[m*]</td>
<td>NUM</td>
<td>_m</td>
<td>five, 5</td>
<td></td>
</tr>
</tbody>
</table>
BYU Corpora: search types

- **List**
  - lists all 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 for seeing change across time

- **Collocates**
  - Shows which words appear in the vicinity of a given word

- **Compare**
  - Can be used to tease apart the different connotations of synonyms

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

- **Text field**
  - 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

- **POS dropdown**
  - 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.)
  - can be used as standalone wildcards or to search for a word as a particular POS
BYU Corpora: Collocates

- **Word/phrase**
  - the main word or phrase you’re searching for

- **Collocates**
  - words that appear in the vicinity of the main word or phrase

- **Numbers**
  - search up to 9 words to the left or right of the main word or phrase
BYU Corpora: Compare

• **Word1 and Word2**
  • the two words you’re comparing

• **Collocates**
  • words that occur in the vicinity of those two words

• **Numbers**
  • search up to 9 words to the left or right of the main word or phrase
BYU Corpora: KWIC (keyword in context)

• Word1 and Word2
  • the two words you’re comparing

• Numbers
  • search up to 9 words to the left or right of the main word or phrase
BYU Corpora: search query syntax

- all inflected forms of a word: put it in brackets or all caps—[word] or WORD
- a word only as a particular part of speech: put a POS tag after it—word.[v*] or 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](#) (or click “About Ngram Viewer” in the footer).
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, e.g. walk_INF_VERB
  - 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.)
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)
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

- Corpora can supplement other reference works and your own native speaker intuition
- 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