Corpus Linguistics

Applied Corpus Search
Corpus of Contemporary American English (COCA)

Niko Schenk

Institut für England- und Amerikastudien
Goethe-Universität Frankfurt am Main
Winter Term 2017/2018

November 29th, 2017

A List of Available Corpora[3]

	Properties			
Corpus	language	words	time period	type
Google's N-Gram Corpus	English	1.024 trillion	-	web data
Google Books Corpus	AE/BE	155/34 billion	1500s-2000s	historical, contemporary books
Global Web-Based English (GloWbE)	20 countries	1.9 billion	2012-2013	web pages
Corpus of Contemporary AE[1] (COCA)	AE	450 million	1990-2012	spoken, fiction, magazines, news, acad texts
British National Corpus (BYU-BNC)[2]	BE	100 million	1980s-1993	representative sample of written/spoken BE
Corpus of American Soap Operas	AE	100 million	2001-2012	film dialogues
Strathy Corpus	Canadian English	50 million	1970s-2000	spoken, fiction, magazines, newspapers, academic texts.
My S-21 Facebook Corpus	German	50 million	2010-2013	UGC, web data
Corpus do Português	Portuguese	45 million	1300s-1900s	newspaper academic texts
Canadian Hansard Corpus	English, French	26 million	1986-1987	parallel corpus, parliament debates
International Corpus of Learner English	English 16 native langs	3.7 million	2002	essays written by learners of English

[1] `http://corpus.byu.edu/coca/`

[2] `https://corpus.byu.edu/bnc/`

[3] no exhaustive list, sorted by size; references: 1, 2

COCA Corpus

Getting started with the **COCA corpus**...
`http://corpus.byu.edu/coca`

Tagset and Instructions on How to Use the Corpus

1. **Tagset**
   `http://ucrel.lancs.ac.uk/claws7tags.html`
2. **Instructions on how to search the data**
   Click on the LIST button and explore **all** links in the section
   More information on: basic syntax, part of speech, lemmata (word base forms), synonyms, customized word lists, etc.

Exercises

Use the **COCA** corpus for your analysis and explore the following exercises. For each exercise,

- provide the query
- provide a short (**brief and concise(!)**) explanation of the trend that you see, based on the obtained frequencies. Note: objective description first, then the interpretation.
- Also note that for some exercises you might want to switch between the display options LIST, CHART, KWIC and COMPARE.

COCA Corpus—Video Lectures

In case you're having trouble with the search or when you need more information on how to work with the corpus you can consult these video lectures:

- About the **COCA** corpus:
  `http://www.youtube.com/watch?v=sCLgRTlxG0Y`
- Parts-of-Speech (POS)
  `http://www.youtube.com/watch?v=KP-7thiUnLM`
- List of POS tags
  `http://ucrel.lancs.ac.uk/claws7tags.html`
- Collocations
  `http://www.youtube.com/watch?v=t_SxpfiPo_o`

Word Meaning

1. Search for the word *corpus*, inspect the results and try to use the different contexts to capture the different meanings.
2. The words *rapid*, *quick*, and *fast* all have the same meaning of *schnell*. Use the corpus to find differences in their use.

Word Frequencies

3. What are the top-five most frequent words in the corpus?
   - What's so special about the second and third most frequent "words"? Why are they included? Think of a potential application/linguistic scenario in which you might want to use these within your search query.
4. What is the most frequent noun in the corpus?
   - Compute the relative frequency of this word compared to all words in the corpus. (simple division)
   - Lookup the same word in the Google NGram viewer `https://books.google.com/ngrams/` and check whether the word's relative frequency in the books corpus is different. Report and compare the two numbers.
5. What are the two most frequent words preceding the word *body*?
   - What are the two most frequent affixes preceding the word *body*? Inspect the results for the seventh most frequent word which looks a bit strange. Could you explain what it is?

Synonyms, POS-Tags, Affixes, Lemmata

6. Find five synonyms of the verb *(to) love*. The synonyms should only be verbs.
7. Click on the keyword-in-context view (KWIC). Search for all nouns of the word form play. Inspect the results and find a sentence which was been tagged incorrectly. (e.g., a sentence in which the word is actually a verb.)
8. What are the three most frequent adjectives starting with the prefix *in*?
   - Restrict your search only to the fiction domain / academic writing genre and report the adjectives.
9. Search for the lemma forms *nice* and *tall* with the LIST display option. Do the same for *good*. What is a potential problem here?
10. *-licious* is a suffix which is used to form new words. Find some instances and come up with a definition for them.

Comparing Genres

11. Are auxiliary verbs used more often in spoken language or in written text?
12. Generally, search for all nouns, verbs, adjectives and adverbs and compare the results across all genres in the corpus. Try to come up with a simple explanation for the trend you see.
13. Formulate a query for passive tense. Show that the *passive tense* is used more often in academic writing compared to fiction texts. What could be a possible explanation?
14. Compare the use of *negation* (not, etc.) and verb (base forms) across genres. (Note, that there is a tag for negation). Explain the trend you see.
15. In fiction texts, you would expect a lot of proper names. How does this hypothesis relate to other genres? Could you think of a linguistic construction (word, part-of-speech tag, ngram, affix) which is more prominent in fiction writing compared to the other genres?

Collocations

16. Search for all adjectives preceding the token *President*. Only inspect the first eleven results. Come up with two linguistic categories for the resulting adjectives by trying to classify them.
17. Which type of nouns does *cause* collocate with?
18. Which type of adjectives does *rather* collocate with? How about *fairly*? Compare the two types of adjectives and inspect many of them carefully. (Use the COMPARE option) Do these two types of adjectives fall into two classes with different properties?
19. Search for *hard* followed by any word. Inspect the results. Then, from the SORTING AND LIMITS panel, choose SORT BY RELEVANCE and rerun the query. Why are the results different? Which one is better interpretable?
20. Which type of nouns follow *handsome*? Which words go with *pretty*? Try to categorize them.

Collocations

21. A guy in a language form[4] claims that "little carries an emotional factor [...] *small* usually does not". Prove this informally.
22. The words *quick*, *rapid* and *fast* all have very similar meanings. Formulate a query which extracts their collocates and explain the differences.
23. The word *them* can (very informally) be used as a synonym for *those*.[5] Find instances of this type in the corpus.

[4] http://www.english-test.net/forum/ftopic14714.html
[5] http://de.urbandictionary.com/define.php?term=them

COCA vs. BNC—Lexicography & Syntax

24. Previous research on quotative like[6] has claimed that the phenomenon is much more common in AE than in BE. Test the hypothesis formally using the corpora COCA and BNC.
25. Formulate a query to check which adjectives are used to describe men. The query should have the pattern *masculine pronoun + form of (to) be* and collocate with adjectives to the right (max 4 tokens). Sort by `RELEVANCE`. Interpret the result. Which of the two lists are you more familiar with?

[6] `http://en.wikipedia.org/wiki/Like#As_a_colloquial_quotative`

COCA vs. BNC—Lexicography & Syntax

26. Compare constructions of the sort *-need NEG VERB-* as in *need not worry* in AE and BE.
27. Search for constructions of the sort *-Beginning of sentence One DO NEGATION-* as in *One doesn't* and compare AE to BE. Could you come up with a hypothesis for the trend you see? (in general/for academic texts?)
28. *-all of the NOUN-* vs. *-all the NOUN-* / *all the cases* vs. *all of the cases* (BNC vs. COCA)
29. Search for all noun collocates of the noun *web*. (4 tokens to the left and right). Compare AE to BE and sort by `RELEVANCE`. Explain the differences.
30. Similar to the previous exercise but with *dumb*.