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 2015/2016

November 30th, 2016
1 COCA Corpus

2 Exercises
COCA Corpus

Exercises
## A List of Available Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google’s N-Gram Corpus</strong></td>
<td>Language: English</td>
</tr>
<tr>
<td></td>
<td>Words: 1.024 trillion</td>
</tr>
<tr>
<td><strong>Google Books Corpus</strong></td>
<td>Time period: 1500s-2000s</td>
</tr>
<tr>
<td>AE/BE</td>
<td>Type: web data</td>
</tr>
<tr>
<td>Global Web-Based English (GloWbE)</td>
<td>20 countries</td>
</tr>
<tr>
<td>1.9 billion</td>
<td></td>
</tr>
<tr>
<td><strong>Corpus of Contemporary AE</strong>&lt;sup&gt;1&lt;/sup&gt; (COCA)</td>
<td>1990-2012</td>
</tr>
<tr>
<td>AE</td>
<td></td>
</tr>
<tr>
<td>British National Corpus (BYU-BNC)</td>
<td>BE</td>
</tr>
<tr>
<td>100 million</td>
<td>Time period: 1980s-1993</td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td></td>
</tr>
<tr>
<td>100 million</td>
<td>Type: film dialogues</td>
</tr>
<tr>
<td>Strathy Corpus</td>
<td></td>
</tr>
<tr>
<td>Canadian English</td>
<td></td>
</tr>
<tr>
<td>50 million</td>
<td></td>
</tr>
<tr>
<td>My S-21 Facebook Corpus</td>
<td>German</td>
</tr>
<tr>
<td>50 million</td>
<td>Time period: 2010-2013</td>
</tr>
<tr>
<td>Corpus do Português</td>
<td>Portuguese</td>
</tr>
<tr>
<td>45 million</td>
<td>Type: newspaper</td>
</tr>
<tr>
<td>Canadian Hansard Corpus</td>
<td>English, French</td>
</tr>
<tr>
<td>26 million</td>
<td>Time period: 1986-1987</td>
</tr>
<tr>
<td>International Corpus of Learner English</td>
<td>16 native langs</td>
</tr>
<tr>
<td>3.7 million</td>
<td>Type: essays written by learners of English</td>
</tr>
</tbody>
</table>

<sup>1</sup> [http://corpus.byu.edu/coca/](http://corpus.byu.edu/coca/)

<sup>2</sup> no exhaustive list, sorted by size; references: 1, 2
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: basic syntax, part of speech, lemmas (forms of words), synonyms, customized word lists, and combining words.
Use the COCA corpus for your analysis and explore the following exercises. For each exercise,

- provide the query that you formulated
- a short (brief and concise(!)) explanation of the trend that you see (based on frequencies that you obtain).
- Also note that for some exercises you might want to switch between the display options List, Chart, KWIC and Compare.
In case you’re having trouble with the search or when you need some 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=sCLgRT1xG0Y](http://www.youtube.com/watch?v=sCLgRT1xG0Y)
- Parts-of-Speech (POS)  
  [http://www.youtube.com/watch?v=KP-7thiUnLM](http://www.youtube.com/watch?v=KP-7thiUnLM)
- List of POS tags  
  [http://ucrel.lancs.ac.uk/claws7tags.html](http://ucrel.lancs.ac.uk/claws7tags.html)
- Collocations  
  [http://www.youtube.com/watch?v=t_SxpfiPo_o](http://www.youtube.com/watch?v=t_SxpfiPo_o)
Search for the word *corpus*, inspect the results and try to use the different contexts to capture the different meanings.
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.

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.

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?
Find five synonyms of the verb *(to) love*. The synonyms should only be verbs.

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.)

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.

Search for the lemma forms *nice* and *tall* with the LIST display option. Do the same for *good*. What is a potential problem here?

*-licious* is a suffix which is used to form new words. Find some instances and come up with a definition for them.
Comparing Genres

10 Are auxiliary verbs used more often in spoken language or in written text?

11 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.

12 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?

13 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.

14 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?
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.

Which type of nouns does *cause* collocate with?

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?

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?

Which type of nouns follow *handsome*? Which words go with *pretty*? Try to categorize them.
20 A guy in a language form claims that “little carries an emotional factor [...] small usually does not”. Prove this informally.

21 The words *quick*, *rapid* and *fast* all have very similar meanings. Formulate a query which extracts their collocates and explain the differences.

22 The word *them* can (very informally) be used as a synonym for *those*. Find instances of this type in the corpus.

---

3http://www.english-test.net/forum/ftopic14714.html
4http://de.urbandictionary.com/define.php?term=them
Previous research on quotative like\textsuperscript{5} has claimed that the phenomenon is much more common in AE than in BE. Test the hypothesis formally using the corpora COCA and BNC.

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?

\textsuperscript{5}http://en.wikipedia.org/wiki/Like#As_a_colloquial_quotative
COCA vs. BNC—Lexicography & Syntax

25. Compare constructions of the sort "need NEG VERB" as in *need not worry* in AE and BE.

26. 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?)

27. "all of the NOUN" vs. "all the NOUN" / *all the cases* vs. *all of the cases* (BNC vs. COCA)

28. 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.

29. Similar to the previous exercise but with *dumb*. 