Corpus resources for word and phrase studies

Will Styler - Fall 2013

Accessing CU's Corpora

- Accessing Babel - `http://verbs.colorado.edu/corpora/logon/logon.html`
  - This is a guide to accessing babel, once you've gotten a user account. If you're interested in getting one, talk to Will or Rebecca.
- Babel Corpus Listing - `http://verbs.colorado.edu/corpora/index.html`
  - This is a listing of all the corpora currently on babel.colorado.edu and available for student use.
- Using grep - `http://tldp.org/LDP/Bash-Beginners-Guide/html/sect_04_02.html`
  - This site discusses the basics of regular expressions, and how to use them. It assumes you're logged in to a computer like babel.
- Grep for Linguists - `http://arts.anu.edu.au/linguistics/misc/comp_resources/grep.html`
  - Yet another site discussing doing corpus searches using grep on a computer like babel.

Finding how words are used

- Google Advanced Search - `http://www.google.com/advanced_search`
  - This is just a slightly more powerful way to google terms.
- The EnronSent Corpus - `http://verbs.colorado.edu/enronsent/`
  - The EnronSent corpus is a special preparation of a portion of the Enron Email Dataset designed specifically for use in Corpus Linguistics and language analysis. This is the best existing source of emails to search through. You'll want to use something like `grep` to search through the corpus.
- Callhome - `http://www.ldc.upenn.edu/Catalog/LDC97S42.html`
  - Callhome is a corpus of transcribed phone calls from every day people to other every day people. Available on Babel.
- BYU-BNC - `http://corpus.byu.edu/bnc/`

  - The BNC (British National Corpus) is a large selection of texts from a variety of sources, searchable online for both word frequencies and word usages.

- The Broadcast News Corpus - `http://www.ldc.upenn.edu/Catalog/LDC97S44.html`
  - The BN corpus is a selection of transcripts from broadcast news shows. Available on Babel.
- The Brown Corpus - `http://en.wikipedia.org/wiki/Brown_Corpus`
  - The Brown Corpus is a selection of 2 million words of English texts, compiled from works published before 1961. Available on Babel.
- Sketch Engine - `sketchengine.co.uk`
  - Click on "IP Auth" from a campus computer. A great tool for looking at word use.

For comparing word frequencies

- Google Ngrams - `http://ngrams.googlelabs.com/`
  - Google Books Ngram viewer lets you find word frequencies across their entire selection of books, by time, and gives a very pretty output.
- CELEX - `http://verbs.colorado.edu/corpora/`
  - CELEX lists frequency, orthography, pronunciations, and a whole bunch of other information about English words. It's the definitive measure for word frequency in English.
- BYU-BNC - `http://corpus.byu.edu/bnc/`
  - The BNC (British National Corpus) is a large selection of texts from a variety of sources, searchable online for both word frequencies and word usages.
- Language Log - Google Frequency Guide `http://itre.cis.upenn.edu/~myl/languagelog/archives/000397.html`
  - A wonderful Language Log post about how to use Google for linguistic research