Google Books Ngram Viewer

P. S. Langeslag

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

N-gram

- Apparently coined 1963
- *Oxford English Dictionary* definition (entry written Sep 2003, revised March 2022):

> A sequence of *n* letters or characters (where *n* is a variable: see N *n.* 6a, 6b), esp. one occurring within a longer sequence such as a passage of text.

Why Use the Concept in Computational Linguistics/Corpus Linguistics?

Why Use the Concept in Computational Linguistics/Corpus Linguistics?

- A “gram” is a word type; the alternative is to lemmatize your corpus,
- And lemmatization is hard;
- But also to facilitate the study of collocations.

(NB Google Books Ngram Viewer does rely on lemmatization.)

What Is Google Books?

What Is Google Books?

- Began in 2002
- Went live in 2004
- Aims to digitize large numbers of books
- Upwards of 25 million books scanned
- Met with a great deal of litigation (notably Author's Guild and the American Association of Publishers)
- The project has slowed down since c. 2012 (but updated corpora came online in 2019!)
- Official (but dated) history page reads "we're not done—not until all of the books in the world can be found by everyone, everywhere, at any time they need them."

What Is the Value of Equipping Google Books with an Ngram Reader?

What Is the Value of Equipping Google Books with an Ngram Reader?

- The largest searchable corpus of print works and ebooks in the history of the world
- Historical value: quantify the historical use of concepts
- Linguistic value: quantify the historical use of words, phrases, spellings
  - Greatly facilitates *OED* attestation research!
- Not feasible to lemmatize so large a corpus *reliably*; bracketing out linguistic entities is the next best approach

Demonstration

books.google.com/ngrams

Terminology

Gram

A sequence of characters

Unigram

A sequence of characters not interrupted by a space (“word”)

Bigram

A sequence of characters interrupted by a single space (“compound”)

Algorithm

Any unigram is scored against the full corpus of unigrams for the chosen language corpus;

Any bigram is scored against the full corpus of bigrams for the chosen language corpus.

Algorithm

Any unigram is scored against the full corpus of unigrams for the chosen language corpus;

Any bigram is scored against the full corpus of bigrams for the chosen language corpus.

Thus a graph plotting a unigram and a bigram is not, strictly speaking, a comparison.

Usage (1/2)

- Enter comma-separated queries to see them plotted against each other
- A wildcard (`*`) returns the top ten matches
  e.g. `the weather is *`
- `gram_INF` returns inflected forms of a lexical form `gram`
  e.g. `seek_INF` returns *sought*, *seek*, *seeking*, *seeks*
- `gram_NOUN`, `gram_VERB`, etc. tries to return only the matching part of speech
  e.g. `feast_VERB` should not find a hit in the sequence "a feast"
- `gram_*` plots all parts of speech for that form against each other
  e.g. `feast_*` returns the noun *feast*, the verb *feast*, the adjective *feast*, and some noise
- Parts of speech on their own return any match
  e.g. `kiss _PRON_ mother` should return "kiss your mother," "kiss my mother," etc., but plotted as a single function;
- Parts of speech preceded by a wildcard are separated out into different matches
  e.g. `kiss *_PRON mother` should return separate statistics on each of "kiss your mother," "kiss my mother," etc.

Usage (2/2)

- Sentence boundaries: `_START_` / `_END_`
- Dependency relations: `weather=>fair,weather=>beautiful,weather=>nice`
- Combined plots: `+`, e.g. `(ale + lager + beer)`
- Subtracted plots: `-`, e.g. `(ale + lager + beer) - (sparkly + sparkly wine + champagne)`
- Divided plots: `/`, e.g. `beer / wine`
- Multiplied plots: `*`, e.g. `fish,(wallaby * 1000)`
- Plots from multiple corpora: `:`, e.g. `wizard:eng_2019,wizard:eng_fiction_2019`
- Syntactic "root": `_ROOT_`, e.g. `_ROOT_=>eat` to return clauses with *eat* as the finite verb

What Are the Limitations of the Google Books Ngram Reader for Linguistic Purposes?

What Are the Limitations of the Google Books Ngram Reader for Linguistic Purposes?

- Skewed corpus (synchronically)
  - Scientific literature overrepresented (e.g. “Figure” vs “figure”)
- Difference in skew over time
  - Early corpus skews towards religion, late corpus towards science
- Disregards print run/readership
- OCR error
  f vs ſ
- Not representative or reliable prior to c. 1800

Bibliography

Younes, Nadja, and Ulf-Dietrich Reips. “Guideline for Improving the Reliability of Google Ngram Studies: Evidence from Religious Terms.” *PLoS ONE* 14 (3 2019). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213554.

Zhang, Sarah. “The Pitfalls of Using Google Ngram to Study Language.” *Wired*, October 12, 2015. https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/.