Google Books N-gram Corpus

- The Google Books Ngram Corpus consists of words and phrases(i.e. ngrams) and their usage frequency over a period of 5 centuries
- 8 languages

Language	#Books	#Tokens
English	4,541,637	468,491,999,592
Spanish	854,649	83,967,471,303
French	792,118	102,174,681,393
German	657,991	64,784,628,286
Russian	591,310	67,137,666,353
Italian	305,763	40,288,810,817
Chinese	302,652	26,859,461,025
Hebrew	70,636	8,172,543,728

- POS tags (automatically generated) and head-modifiers
- The data is available for download (up to 5-grams), and can also be viewed through the interactive online Google Books Ngram Viewer at `http://books.google.com/ngrams`

Creation

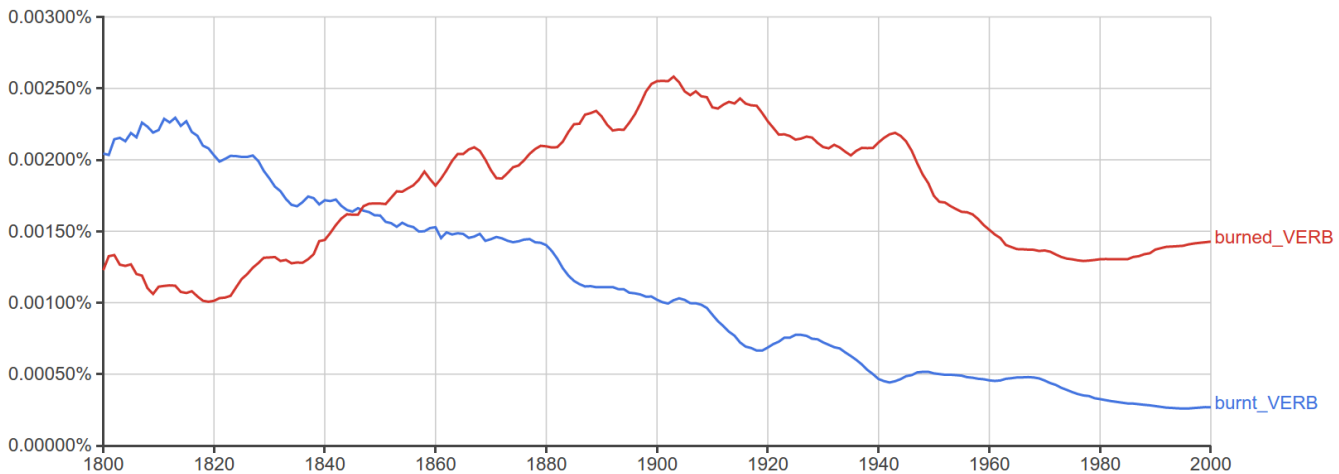
- The Google Books Ngram corpus contains over 8 million books, or over 6% of all books ever published
- This collection of books is much larger than any other digitized collection
- Generation of corpus required a substantial effort involving obtaining and manually scanning millions of books
- Only ngrams that appear over 40 times across the corpus are included
- **File format:**
  `ngram TAB year TAB match_count TAB volume_count NL`
  Example: circumvallate 1978 335 91

Annotation

- 12 universal POS tagset were used:
  N_{OUN} (nouns), V_{ERB} (verbs), A_{DJ} (adjectives), A_{DV} (adverbs), P_{RON} (pronouns), D_{ET} (determiners and articles), A_{DP} (prepositions and postpositions), N_{UM} (numerals),C_{ONJ} (conjunctions), P_{RT} (particles), '.' (punctuation marks) and X (other categories)
- Dependency relation are shown by arrows from head word to the modifier word (e.g., $head => modifier$ or $modifier <= head$)
- Files are classified by n-grams (1-gram, 2-gram,..., 5-gram) and each file name consists of 2 symbols (which denote first 2 symbols of words (n-grams) included in the file)

Usage

- Enables the quantitative analysis of linguistic and cultural trends as reflected in millions of books over the past 5 centuries
- Burned vs Burnt



- `Subject-verb-object` triplet for describing short videos

Reference

- Y. Lim, J. Michel, E. Aiden et al. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 169-174*. (`http://dl.acm.org/citation.cfm?id=2390499`)
- N. Krishnamoorthy, G. Malkarnenkar, R. Mooney et al. (2013). Generating natural-language video descriptions using text-mined knowledge. *Proceedings of the Workshop NAACL HLT, pages 10-19.*

ISLRN

Title	Google Books Ngram Corpus
Full title	Google Books Ngram Corpus
Resource type	Text
Source/URL	https://books.google.com/ngrams/info
Format/MIME type	Text/xml
Size/Duration	863.4GB (7 Aug 2012)
Access medium	online
Description	The Google Books Ngram Corpus describes how often words and phrases(n-grams) were used over a period of 5 centuries, in 8 languages; it reflects 6% of all books ever published. The English corpus alone comprises close to half a trillion words. Words are tagged with their part-of-speech and head modifier relationships. Only ngrams that appear over 40 times across the corpus are included.
Version	2
Media type	Text
Languages	(eng)English, (spa)Spanish, (rus)Russian, (ita)Italian, (heb)Hebrew, (fre)French, (ger)German, (cmn)Chinese
Resource creator	Google Inc.
Distributor	Google Inc.
Rights holder	Google Inc.