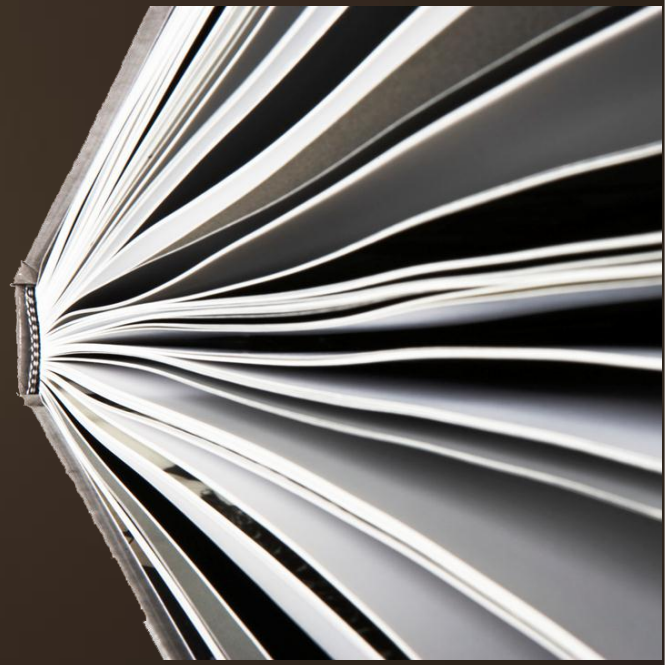
Exploring the Google Books Ngram Viewer for "Big Data" Text Corpus Visualizations

SHALIN HAI-JEW
KANSAS STATE UNIVERSITY
SIDLIT 2014 (OF C2C)
JULY 31 – AUG. 1, 2014



Presentation Overview

- As part of the Google Books digitization project, the Google Books Ngram Viewer (https://books.google.com/ngrams) was released in late 2010 to enable public querying of a "shadow dataset" created from the tens of millions of digitized books. The texts are from a 500-year span (1500-2000+), with new texts added fairly continuously, and there are a range of datasets of different text corpuses (and in different languages, like Italian, French, German, Spanish, Russian, Hebrew, and simplified Chinese). The name of the tool comes from a computer science term referring to strings of alphanumeric terms in particular order: a unigram (or one-gram) consists of one entity, a bigram (or two-gram) consists of two entities, onwards. (Its precursor was a prototype named "Bookworm.") Users may acquire the (de-contextualized) word or phrase or symbol frequency counts of terms in books—which provide a lagging indicator of trends (over time), public opinion, and other phenomena. The Ngram Viewer has been used to provide insights on diverse topics such as the phenomena of fame (and the fields which promote fame), collective forgetting, language usage, cultural phenomena, technological innovations, and other insights. The data queries that may be made with this tool are virtually unanswerable otherwise. The enablements of the Google Books Ngram Viewer provide complementary information sourcing for designed research questions as well as free-form discovery. This tool is also used for witty data visualizations (such as simultaneous queries of "chicken" and "egg" to see which came first) based on the resulting plotted line chart. The tool also enables the download of raw dataset information of the respective ngrams, and the findings are released under a generous intellectual property policy. This presentation will introduce this semi-controversial tool and some of its creative applications in research and learning.


Viewer
Ngram
Google
Books
Research
data
language
English
time
University
may
corpuses
languages
analysis
terms
texts
cont
datasets
information
books
years
tool
Search
research
dataset
Understandings
Russian
Spanish
Chinese
frequency
operator
insights
counts
human
phenomena
n-grams
particular
extractions
used
others
Case
word
various
knowledge
German
Data
visualizations
different
Beijing:eng
linegraph
Bookworm
multilingual
Big
speech
making
INF
VERB
number
learning
query
queries
big
literacy
simplified
complementary
predominantly
technological
ride
French
Italian
innovations
History
term
understandings
text
Overview
One
Knowing
discovery
present
Work
Using
HUMINT
SIGINT
OSINT
MASINT
TECHINT
CYBINT
released
people
collective
potential
Insensitive
Sensitive
Applications
provide
verb
Hebrew
NOUN
phrase
Inflection
DNINT
START
words
brain
count
first
Tips
also
ROOT
machine
Approaches
Two
buy
corpus
average
Text
vs
Searches
Collections
changing
fame
play
sim
main
much
Shadow
underlying
computational

Welcome!

- Hello! Who are you?
- Any direct experiences with n-grams? Any research angles that may be informed by n-grams?
- What are your interests in terms of the Google Books Ngram Viewer? What is your level of experience with this Viewer?

What is the Google Books Ngram Viewer?

A SIMPLE OVERVIEW

History

- Google Books project (conceptualized from 1996, official secret launch in 2002, announcement of "Google Print" project in 2005, new user interface in 2007, known also as Google Book Search)
  - Over 100 million digitized books from 1500s to the present
  - Book Search interface available in 35 languages
  - Over 10,000 publishers and authors from 100+ countries in the Book Search Partner Program
  - Integrated with Google Web Search
  - Public domain works in full view, copyrighted works with snippets ("Google Books")

History (cont.)

- Derived shadow dataset: Bookworm Ngrams -> Ngram Viewer
  - Based on a "bag of words" approach
  - Launched in late 2010
- Google Books Ngram Viewer prototype (then known as "Bookworm") created by Jean-Baptiste Michel, Erez Aiden, and Yuan Shen…and then engineered further by The Google Ngram Viewer Team (of Google Research)

History (cont.)

- Includes a number of corpuses across many languages (finer details of each corpus)
- Current corpuses
  - American English 2012, American English 2009
  - British English 2012, British English 2009
  - Chinese 2012, Chinese 2009
  - English 2012, English 2009
  - English Fiction 2012, English Fiction 2009
  - English One Million
  - French 2012, French 2009
  - German 2012, German 2009
  - Hebrew 2012, Hebrew 2009
  - Spanish 2012, Spanish 2009
  - Russian 2012, Russian 2009
  - Italian 2012

Some Terminology

- **Digital humanities:** Research from computation and disciplines in the humanities
- **Big data:** Datasets with an "n of all," large datasets with a large number of records (such as in the millions)
- **N-gram:** A contiguous sequence of n items from text or speech (unigram, bigram / digram, trigram, four-gram, five-gram, and so on), often representing a concept
- **Text corpuses:** Collections of text such as manuscripts or microblogging streams or other texts
- **Shadow dataset**: Masked or de-identified extrapolated data

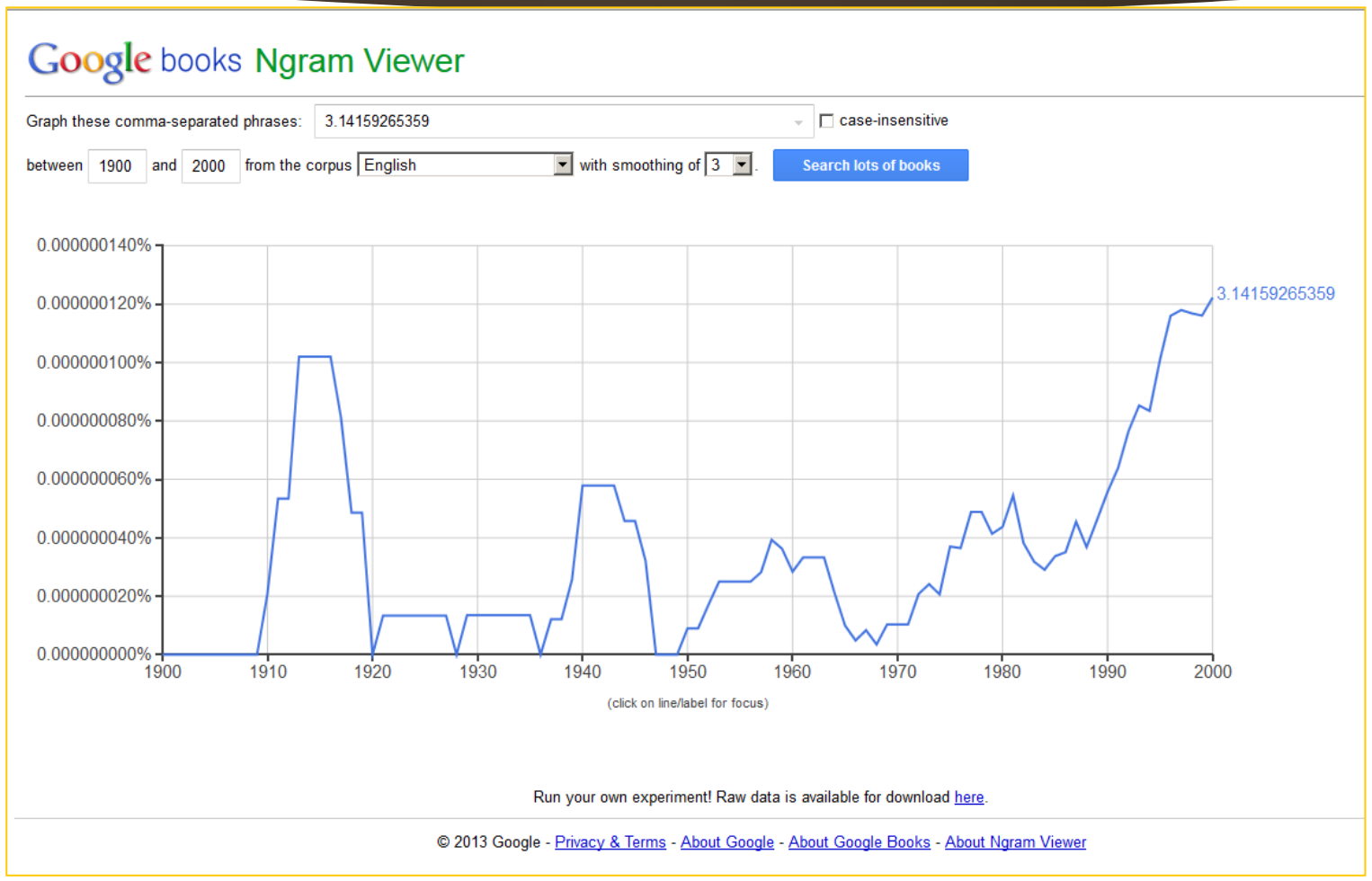
Some Terminology (cont.)

- **Frequency count**: The number of times a particular n-gram appears in a text corpus
- **Smoothing**: The softening of spikes by averaging data from preceding and following years to indicate a "moving average" (a smoothing of 1 means a datapoint on the linegraph is the average of the count for 1 year to each side; a smoothing of 2 means the average of the count for 2 years to each side, etc.)
- **Data visualization**: The image-based depiction of data for the identification of patterns

N-grams

Number of Ngrams	Examples (comma-separated)
Unigram or one-gram	time, $, Julia, pi, 3.14159265359
Bigram or digram or two-gram	borrow money, return home, déjà vu, golden mean
Trigram	her own purse, the trip abroad
Four-gram	the time it took, the Merchant of Venice
Five-gram	when he left the store, after she fed the dog
Six-gram	I plan to travel very soon
Seven-gram	The president will visit the state tomorrow

<3.14159265359>



A "Shadow Dataset" of Google Books Collections

- Shadowing the dataset to protect against privacy infringement or reverse engineering of manuscripts
  - The machine extraction of n-grams from the de-contextualized texts
  - Pure frequency counts of the n-grams in various sequences (but only if they reach a threshold of ngrams that appear in 40 or more books to keep the processing manageable)
  - Tagging of parts of speech (POS) that structure language but do not hold semantic value
  - The elimination of unique phrase sequences to avoid potential hacking and reverse-engineering of particular texts

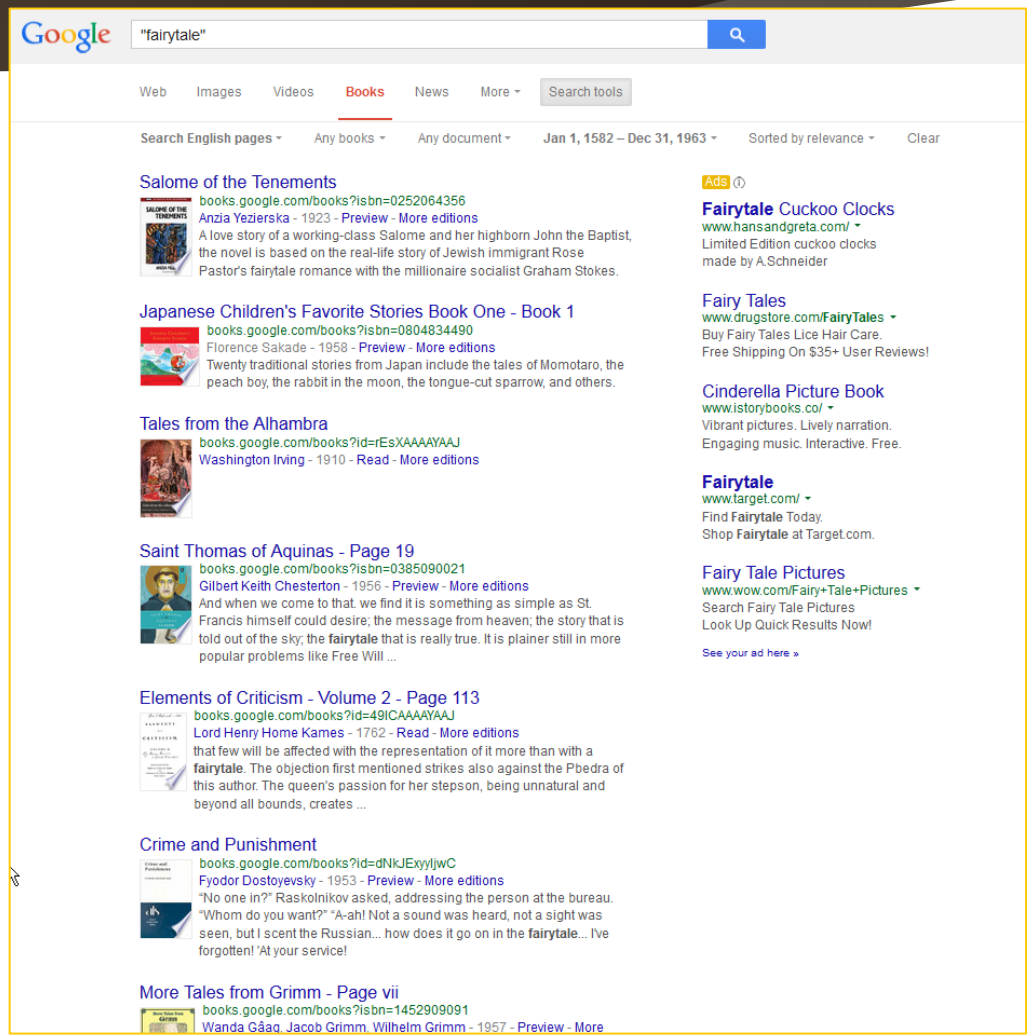
A "Shadow Dataset" of Google Books Collections (cont.)

- Depiction of frequency counts over time (with defined and editable start-and-end years) for broad-scale trending
- Ability to compare multiple words and phrases

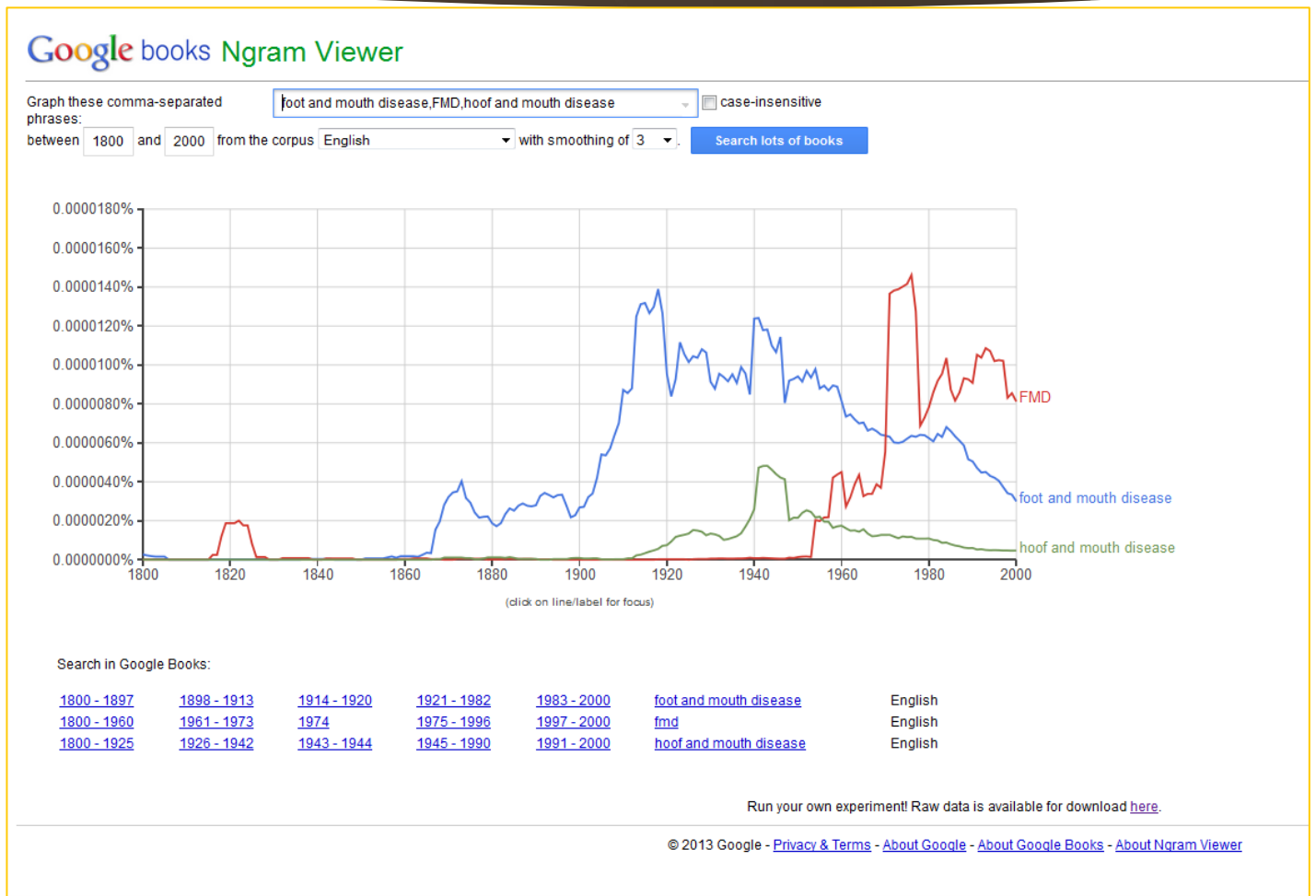
Value Added Capabilities

- Downloadable n-gram datasets (for further analysis)
- Interactive visualizations from mouseovers
- Machine-highlighted years of interest
- Linkage to original texts (on Google)
- Choices from dozens of different and multilingual corpuses (French, simplified Chinese, Italian, Russian, Spanish, Hebrew, German, and others)

Anomalous Years of Interest; Links



<foot and mouth disease, FMD, hoof and mouth disease>



Downloadable Experimental Datasets

The format of the `total_counts` files are similar, except that the `ngram` field is absent and there is one triplet of values (match_count, page_count, volume_count) per year.

Usage: This compilation is licensed under a Creative Commons Attribution 3.0 Unported License.

English

Version 20120701

total_counts

1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

2-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

3-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

4-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

5-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dq dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_ ga gb gc gd ge gf gg gh gi gj gk gl gm gn go gp gq gr gs gt gu gv gw gx gy gz h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jb jc jd je jf jg jh ji jj jk jl jm jn jo jp jq jr js jt ju jv jw jx jy jz k_ ka kb kc kd ke kf kg kh ki kj kk kl km kn ko kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn mo mp mq mr ms mt mu mv mw mx my mz n_ na nb nc nd ne nf ng nh ni nj nk nl nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy oz p_ pa pb pc pd pe pf pg ph pi pj pk pl pm pn po pp pq pr ps pt pu punctuation pv pw px py pz q_ qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs rt ru rv rw rx ry rz s_ sa sb sc sd se sf sg sh si sj sk sl sm sn so sp sq sr ss st su sv sw sx sy sz t_ ta tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv uw ux uy uz v_ va vb vc vd ve vf vg vh vi vj vk vl vm vn vo vp vq vr vs vt vu vv vw vx vy vz w_ wa wb wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo xp xq xr xs xt xu xv xw xx xy xz y_ ya yb yc yd ye yf yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

dependencies 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a b c d e f g h i j k l m n o other p punctuation q r s t u v w x y z

Version 20090715

total_counts

1-grams 0 1 2 3 4 5 6 7 8 9

2-grams 0 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 51 52 53 54 55 56 57 58 59 6 60 61 62 63 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78 79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

3-grams 0 1 10 100 101 102 103 104 105 106 107 108 109 11 110 111 112 113 114 115 116 117 118 119 12 120 121 122 123 124 125 126 127 128 129 13 130 131 132 133 134 135 136 137 138 139 14 140 141 142 143 144 145 146 147 148 149 15 150 151 152 153 154 155 156 157 158 159 16 160 161 162 163 164 165 166 167 168 169 17 170 171 172 173 174 175 176 177 178 179 18 180 181 182 183 184 185 186 187 188 189 19 190 191 192 193 194 195 196 197 198 199 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 43 44 45 46 47 48 49 5 50 51 52 53 54 55 56 57 58 59 6 60 61 62 63 64 65 66 67 68 69 7 70 71 72 73 74 75 76 77 78 79 8 80 81 82 83 84 85 86 87 88 89 9 90 91 92 93 94 95 96 97 98 99

Some Controversies with the Ngram Viewer

- Decontextualized machine "analysis" vs. contextualized reading and human expertise
- Machine "(non)reading" (in frequency counts) vs. human reading (symbolic decoding), a quantitative vs. a qualitative focus, an overbalance into computational understandings (a quantity of words separated from conscious expressed meaning and author hand)
  - Example from Karen Reimer's "Legendary, Lexical, Loquacious Love" (1996), a deconstructed book which consisted of lists of alphabetized contents (per *Uncharted…)*
- Inability to verify results outside of Google Books Ngram Viewer
  - Some degree of elusive "black box" lack of knowledge about functionality

Google books Ngram Viewer

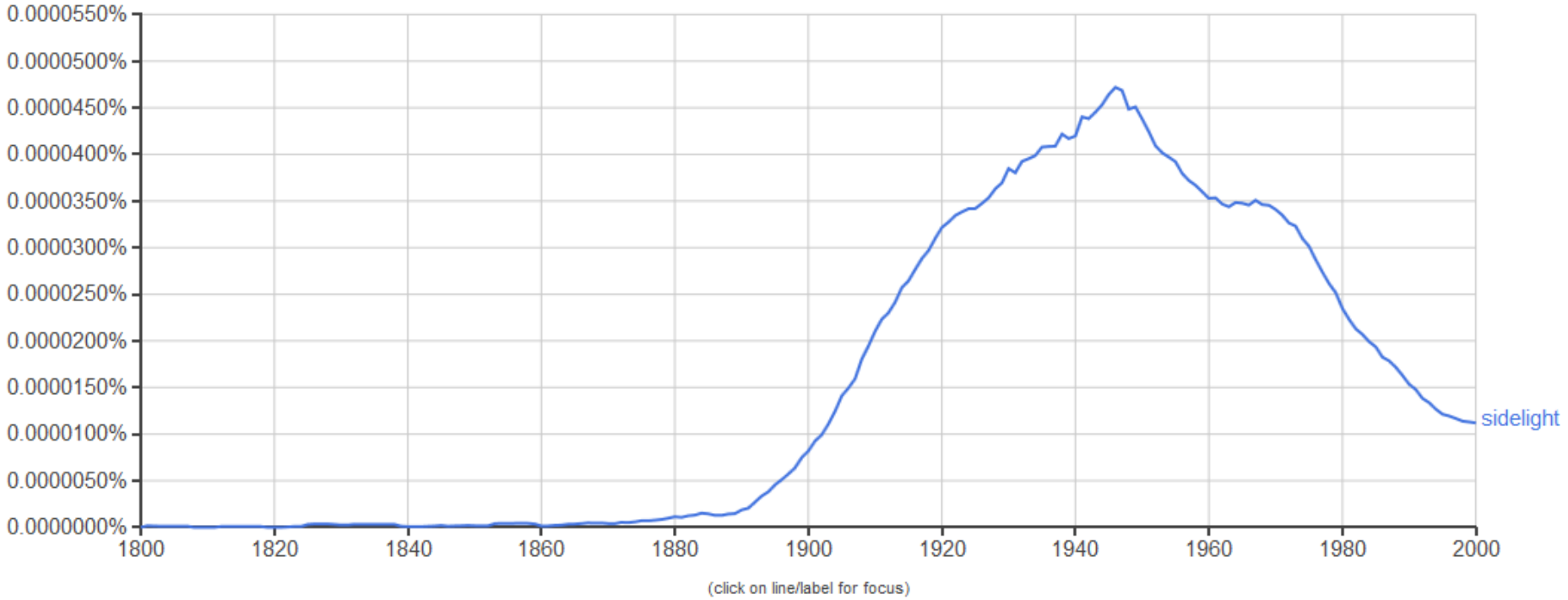
Graph these comma-separated phrases: sidelight case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books

Share 0

Tweet

Embed Chart



(click on line/label for focus)

Search in Google Books:

1800 - 1913 1914 - 1945 1946 - 1952 1953 - 1979 1980 - 2000 sidelight English

Run your own experiment! Raw data is available for download here.

Research Potential?

- On first blush, what do you think can be learned from such data extractions? Why?
- What can be asserted from the linegraphs?
- Would publishers ever accept a deconstructed book for publication (or do you think these are mainly one-offs??

A Cursory Overview of Research Findings So Far from Ngram Viewer

- **Fame**: Who gets famous, and how? What sorts of professions lead to fame?
- **Collective memory:** In terms of human memories of events, how long do people tend to remember? What is the trajectory of collective consciousness from knowing to not knowing?
- **Adoption of innovations:** What is the typical time length for people to accept technological and other innovations? How do cultural phenomena affect human populations over time?
- **Language evolution:** How does language evolve over time? How do rules of language become normalized?

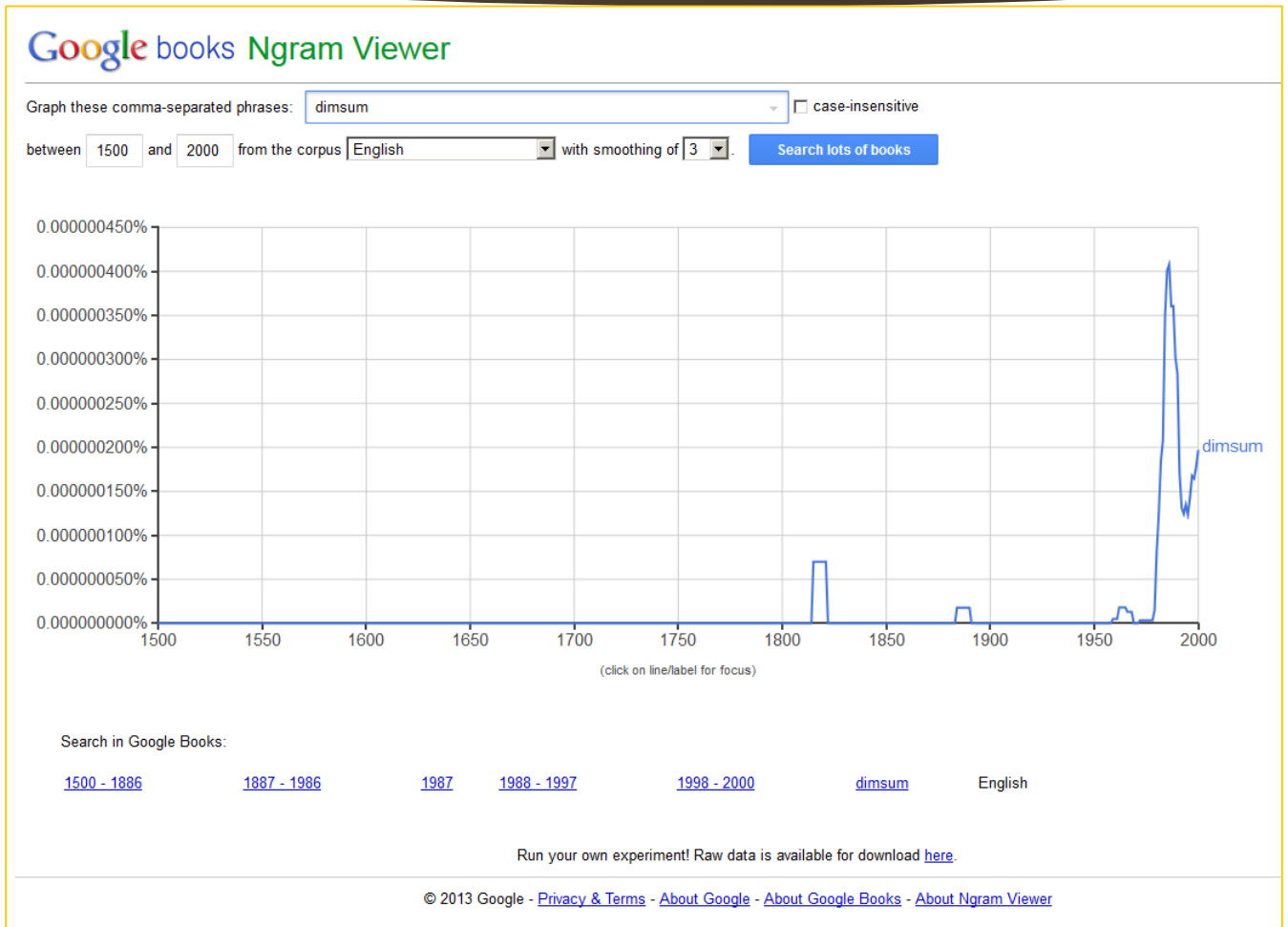
A Cursory Overview of Research Findings So Far from Ngram Viewer (cont.)

- **First-use of terms:** When was the time when a term was first used? (such as terminology linked to technological innovations) (a form of fact-checking)
- **Popularity**: Between various artists / scientists / politicians, who was more popular in his / her day?
- **Government Censorship**: What was the role of Nazi censorship of certain Jewish artists in terms of their reputations and mentions / non-mentions in the literature?

Some Examples

USING GENERAL DATA EXTRACTIONS

Single Extraction <dimsum>



Two Element Comparison <dimsum,tapas>

Two Element Comparison (cont.) <future,past>

Multi-Element Comparisons and Contrasts <height,weight> <diet,exercise>

Google books Ngram Viewer

Graph these comma-separated phrases: height,weight case-insensitive

between 1500 and 2000 from the corpus English with smoothing of 3. Search lots of books

Search in Google Books:

1500 - 1637	1638 - 1666	1667 - 1811	1812 - 1948	1949 - 2000	height	English
1500 - 1640	1641 - 1672	1673 - 1818	1819 - 1961	1962 - 2000	weight	English

Run your own experiment! Raw data is available for download here.

Google books Ngram Viewer

Graph these comma-separated phrases: diet,exercise case-insensitive

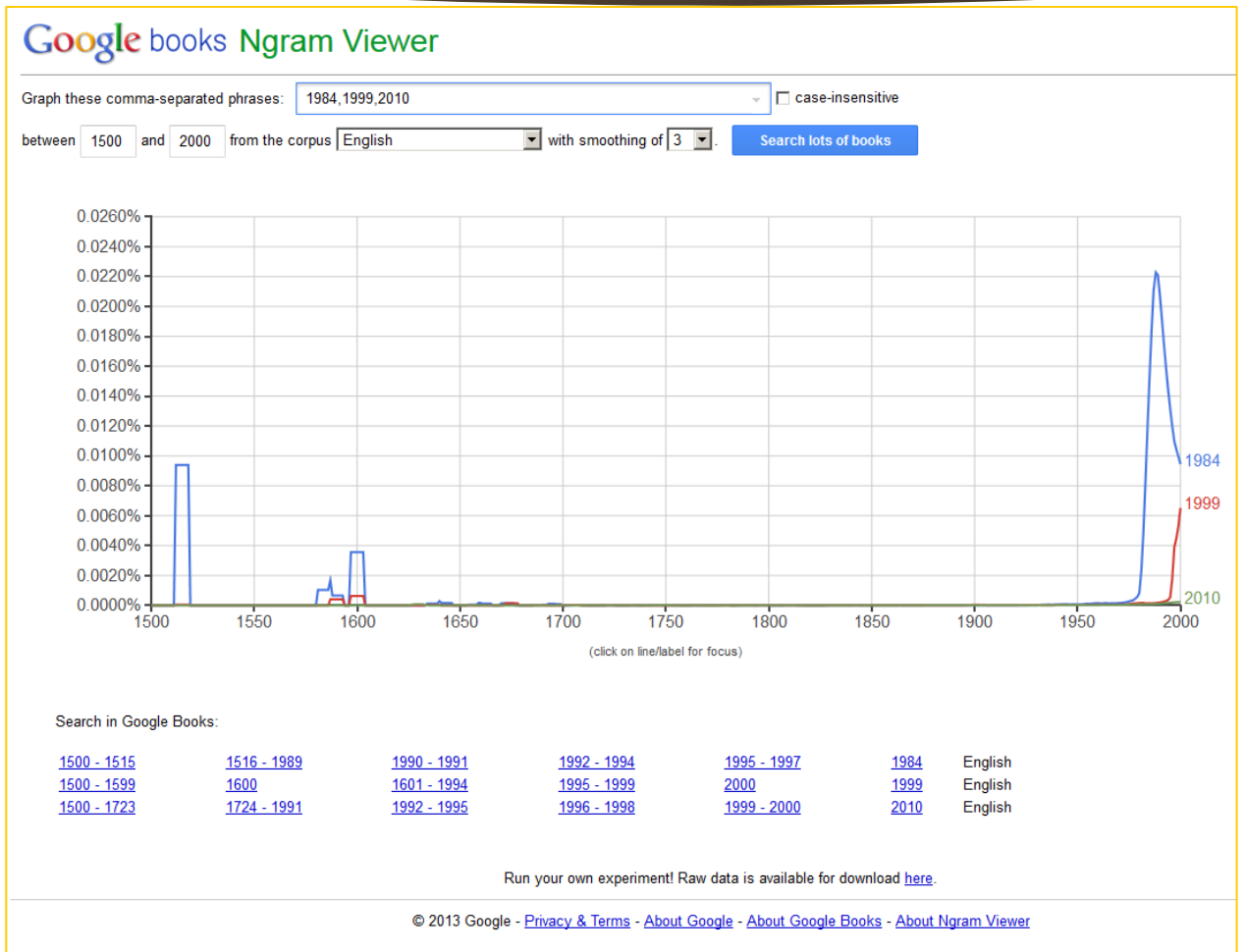
between 1500 and 2000 from the corpus English with smoothing of 3. Search lots of books

Search in Google Books:

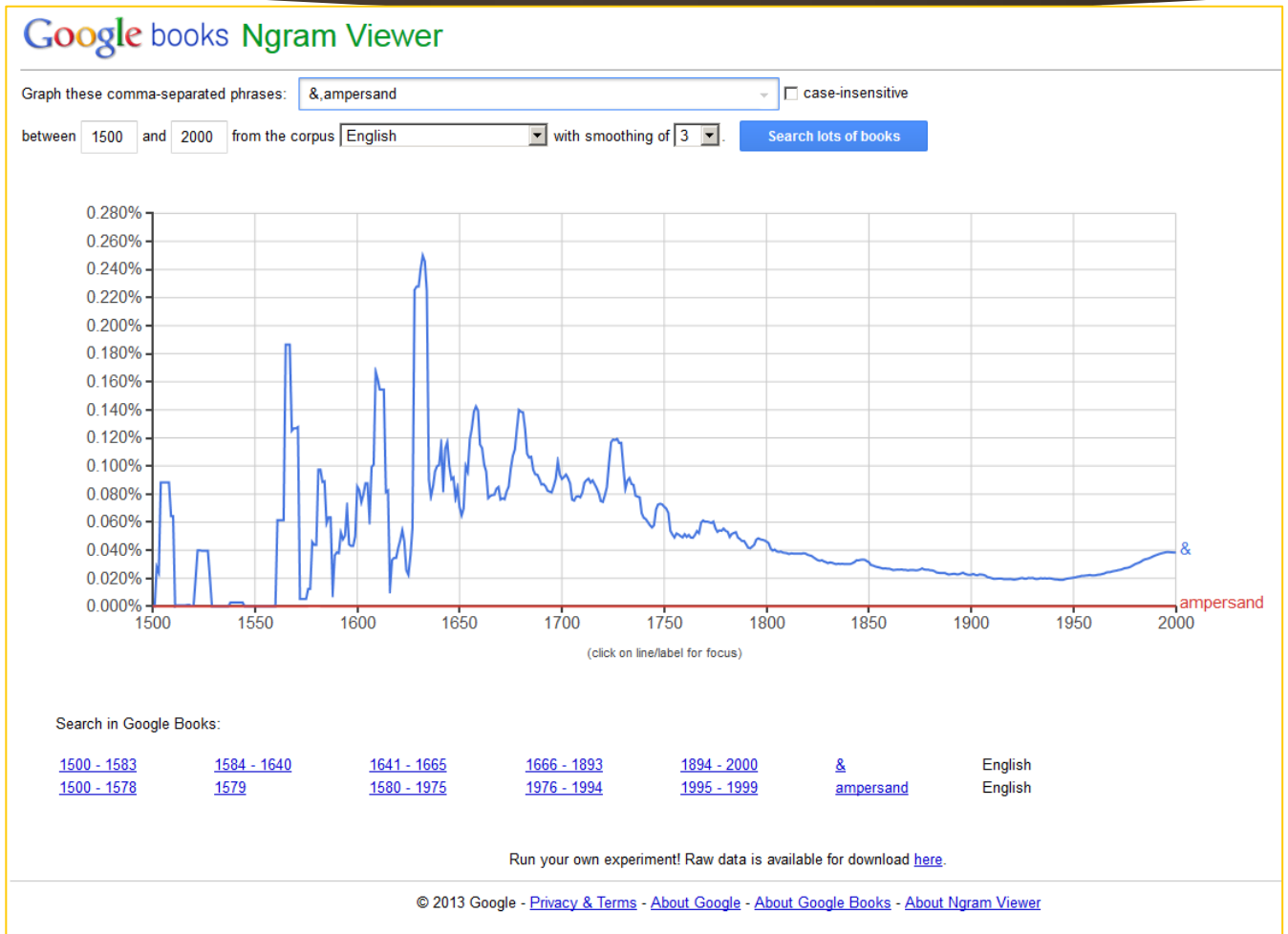
1500 - 1629	1630 - 1773	1774 - 1798	1799 - 1973	1974 - 2000	diet	English
1500 - 1620	1621 - 1806	1807 - 1836	1837 - 1956	1957 - 2000	exercise	English

Run your own experiment! Raw data is available for download here.

Year(s) <1984, 1999, 2010>



Symbols
<&, ampersand>



Phrases: Multiword Strings <crossing the Rubicon>

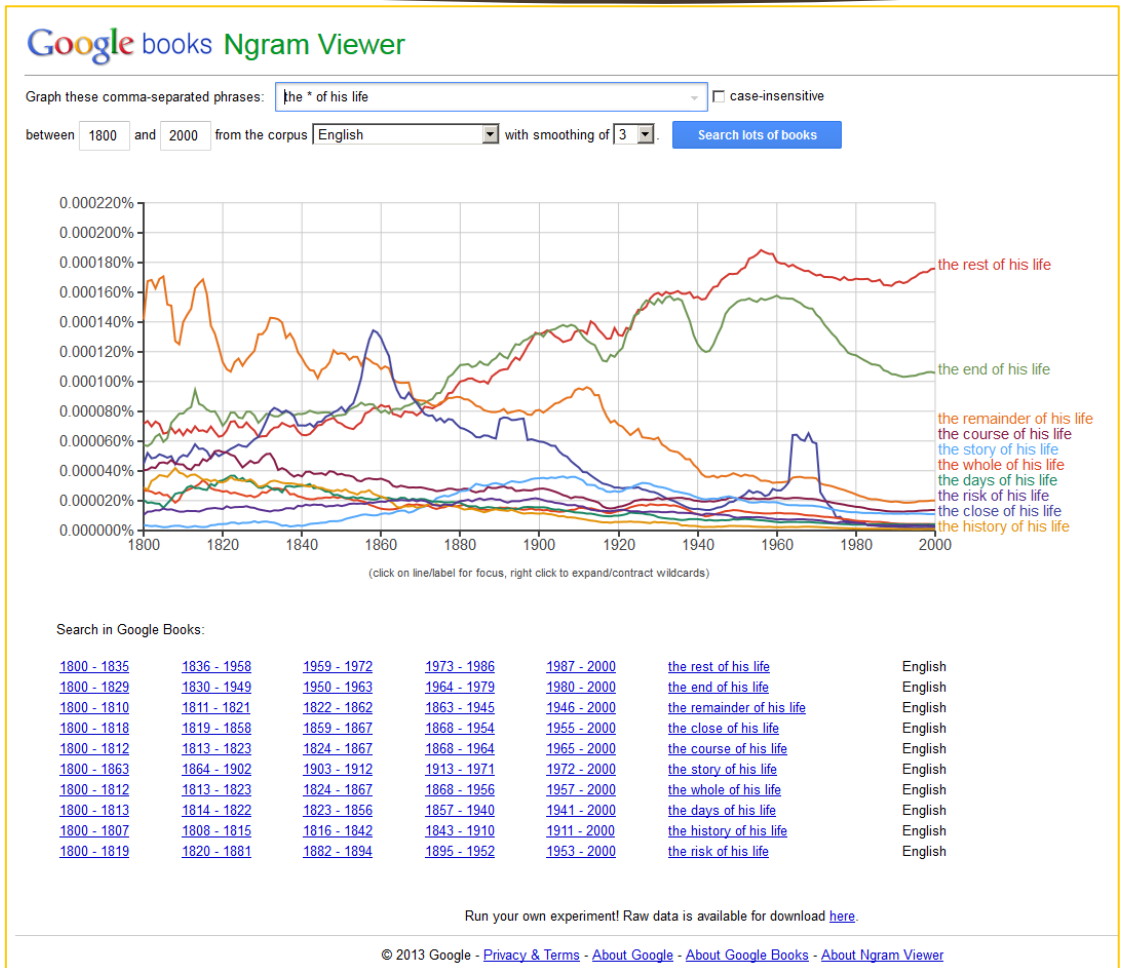
(Somewhat More) Advanced Searches

- WILDCARD EXTRACTIONS
- INFLECTION SEARCHES
- CASE SENSITIVITY / CASE INSENSITIVITY
- ACCESSING TAGGING (AS FOR PARTS OF SPEECH)
- RICHER COMBINATIONS
- SOME BOOLEAN-BASED QUERIES
- STARTS AND ENDS OF SENTENCES
- DEPENDENCY RELATIONS
- ROOTS OF PARSE TREES

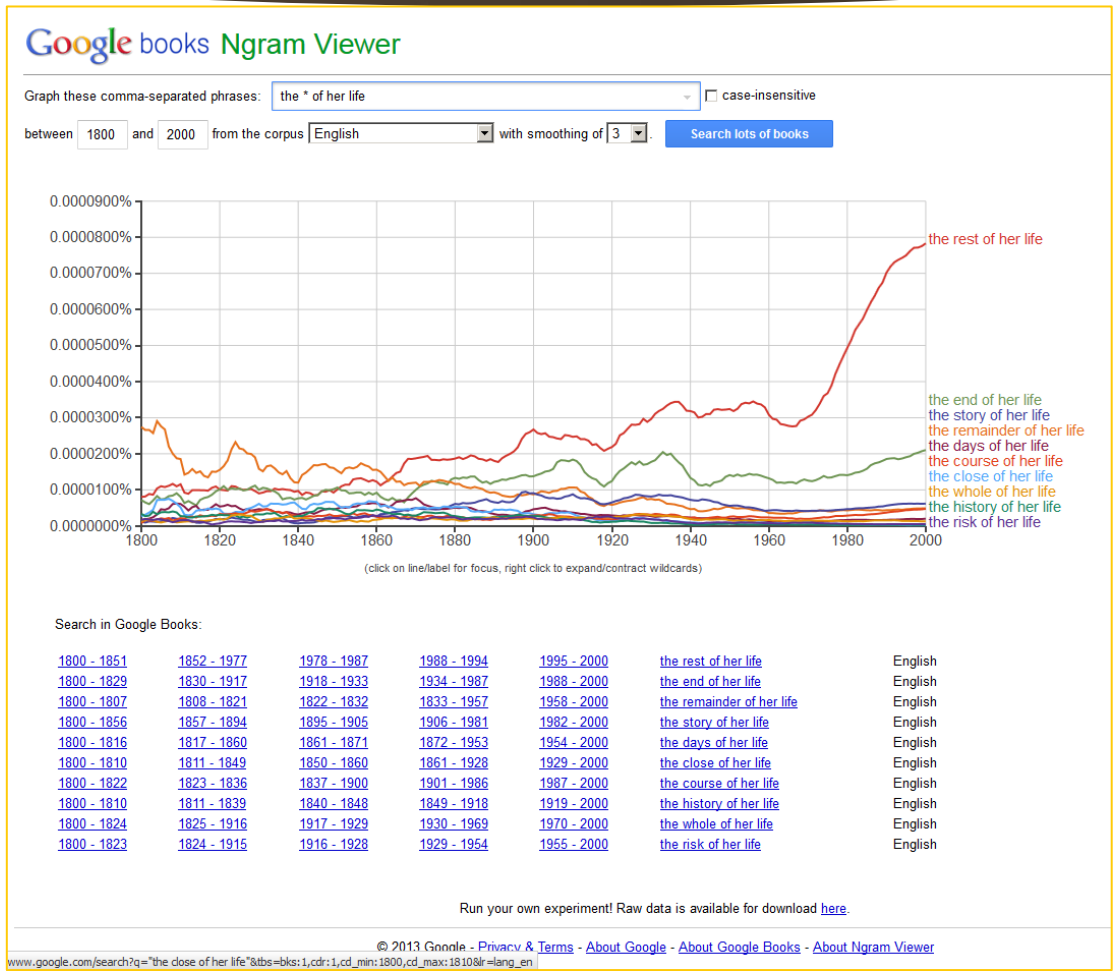
Wildcard Search *

- The use of * to stand in for a word so the Ngram Viewer will display the top ten substitutes (of most common stand-in words) for the asterisk

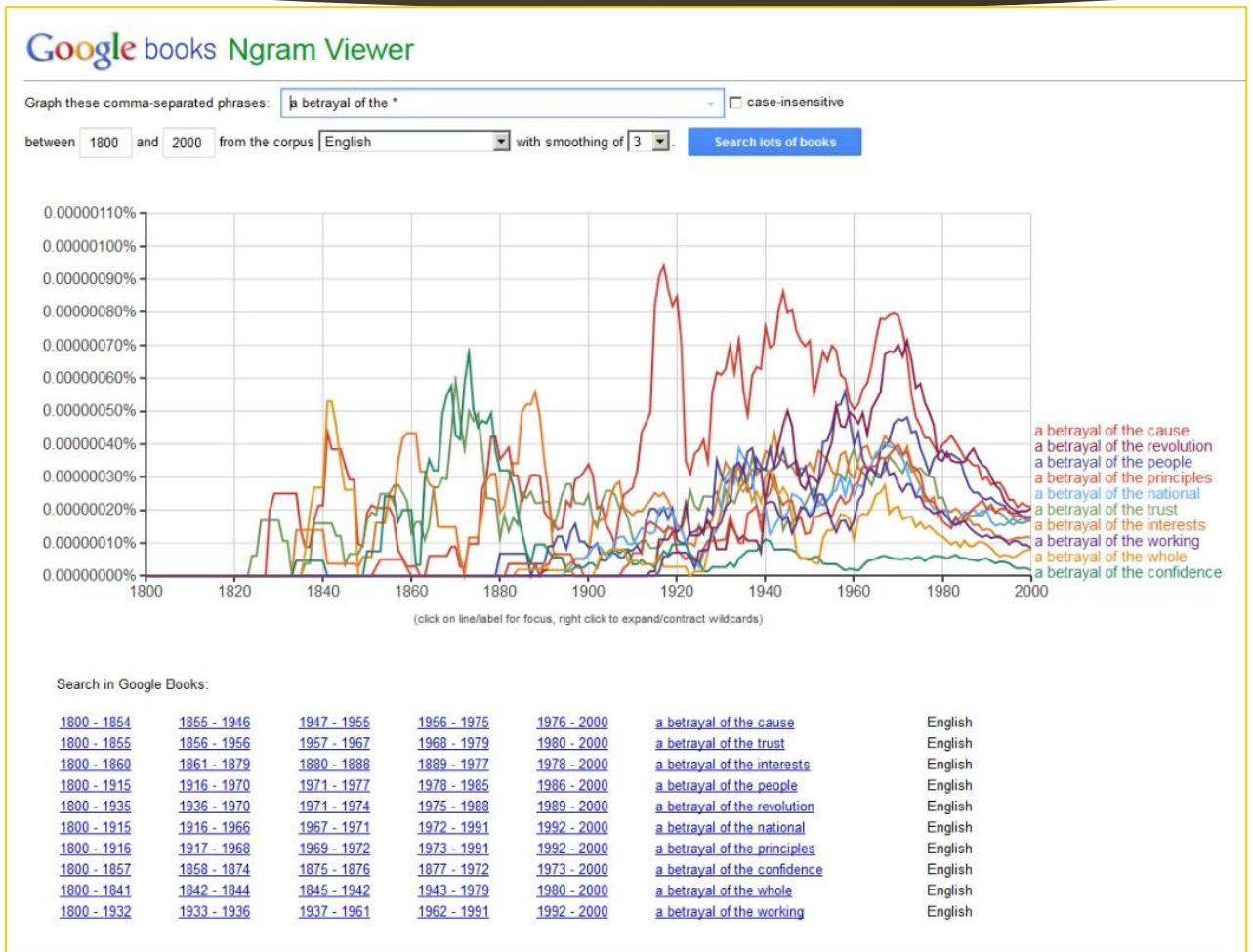
<the * of his life>



<the * of her life>



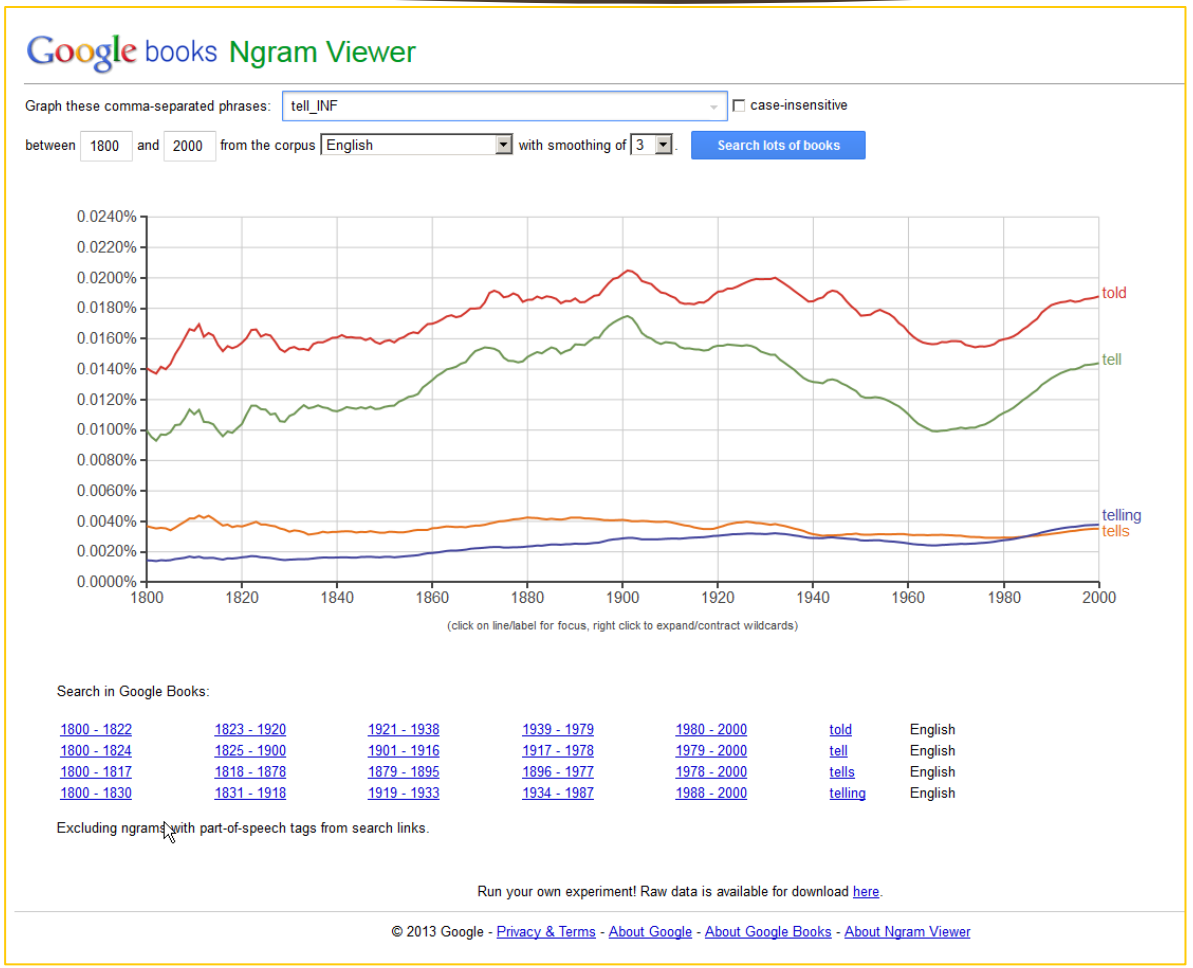
<a betrayal of the *>



Inflection Search with _INF

- Differentiation of various word forms
  - (infinite verb form)
  - -ed
  - -ing
  - -s
  - (irregular spellings)

<tell_INF>


Google books Ngram Viewer
Graph these comma-separated phrases: tell_INF
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3
Search lots of books
0.0240%
0.0220%
0.0200%
0.0180%
0.0160%
0.0140%
0.0120%
0.0100%
0.0080%
0.0060%
0.0040%
0.0020%
0.0000%
1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000
told
tell
telling
tells
(click on line/label for focus, right click to expand/contract wildcards)
Search in Google Books:
1800 - 1822 1823 - 1920 1921 - 1938 1939 - 1979 1980 - 2000 told English
1800 - 1824 1825 - 1900 1901 - 1916 1917 - 1978 1979 - 2000 tell English
1800 - 1817 1818 - 1878 1879 - 1895 1896 - 1977 1978 - 2000 tells English
1800 - 1830 1831 - 1918 1919 - 1933 1934 - 1987 1988 - 2000 telling English
Excluding ngrams with part-of-speech tags from search links.
Run your own experiment! Raw data is available for download here.
© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

Case Sensitive / Insensitive Searches

Case Sensitive

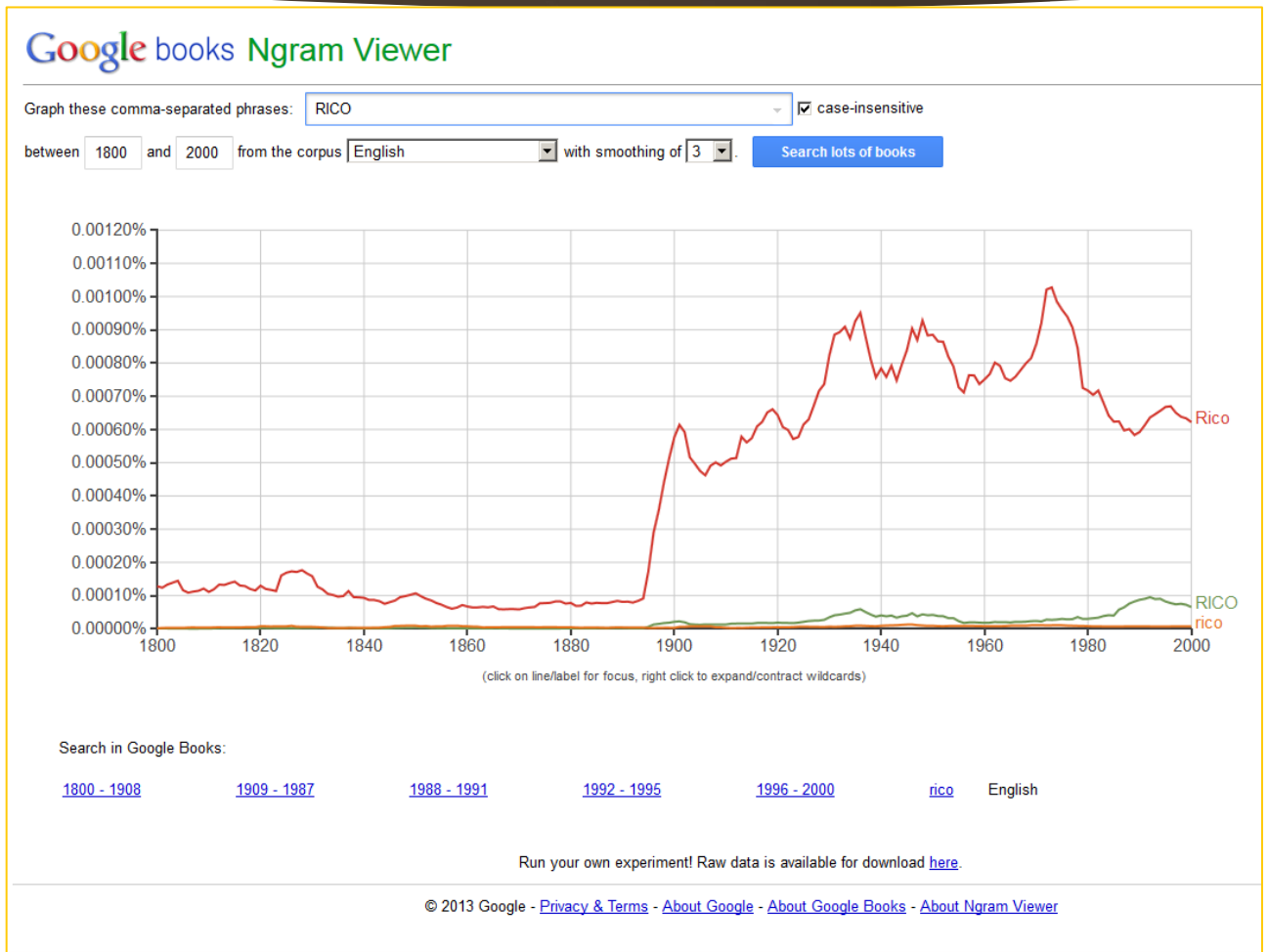
- Capitalizations and lower cases matter

Case Insensitive

- Capitalizations and lower cases do not matter
- Case insensitivity will result in a variety of capitalization / lower case mixes and variations for a particular search term

Case Sensitive <RICO>

Case Insensitive <RICO>

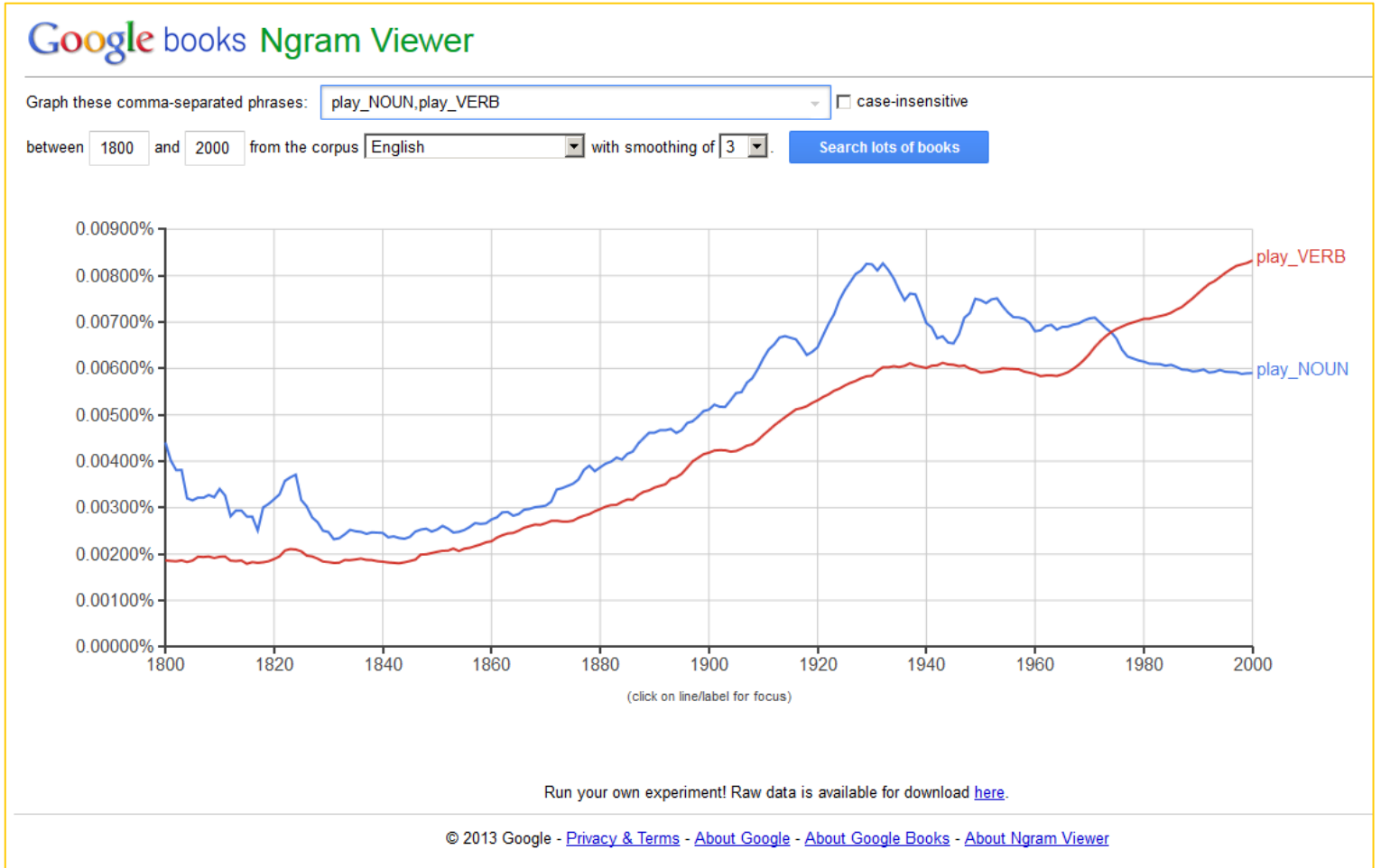


Part-of-Speech Tags

- Disambiguation of term to defining its usage as a part-of-speech to capture the conceptual usage
- May be used as stand-alones (_VERB_) or appended to a verb (play_VERB)

TAGS	APPLICATION
NOUN	Noun
VERB	Verb
ADJ	Adjective
ADV	Adverb
PRON	Pronoun
DET	Determiner or article
ADP	Adposition (preposition or postposition)
NUM	Numeral
CONJ	Conjunction
PRT	Particle
ROOT	Root of the parse tree
START	Start of a sentence (sentence boundary)
END	End of a sentence (sentence boundary)

<play_NOUN, play_VERB>



Some Combinations

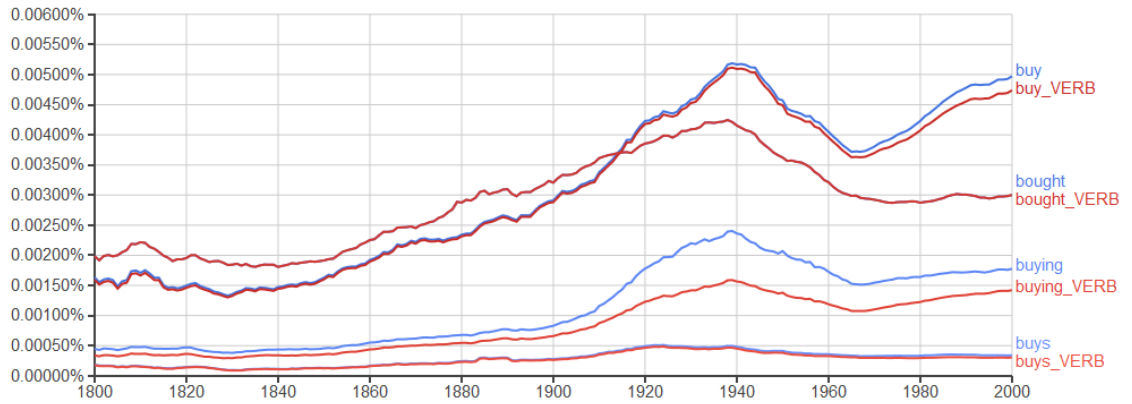
- Inflection keyword with part-of-speech text
  - buy_INF, buy_VERB_INF (buy, buying, bought, buys)
- Dependencies with wildcards
  - ride=>*_NOUN (ride car; ride bike; ride bus)

<buy_INF, buy_VERB_INF>

Google books Ngram Viewer

Graph these comma-separated phrases: buy_INF,buy_VERB_INF case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books



(click on line/label for focus, right click to expand/contract wildcards)

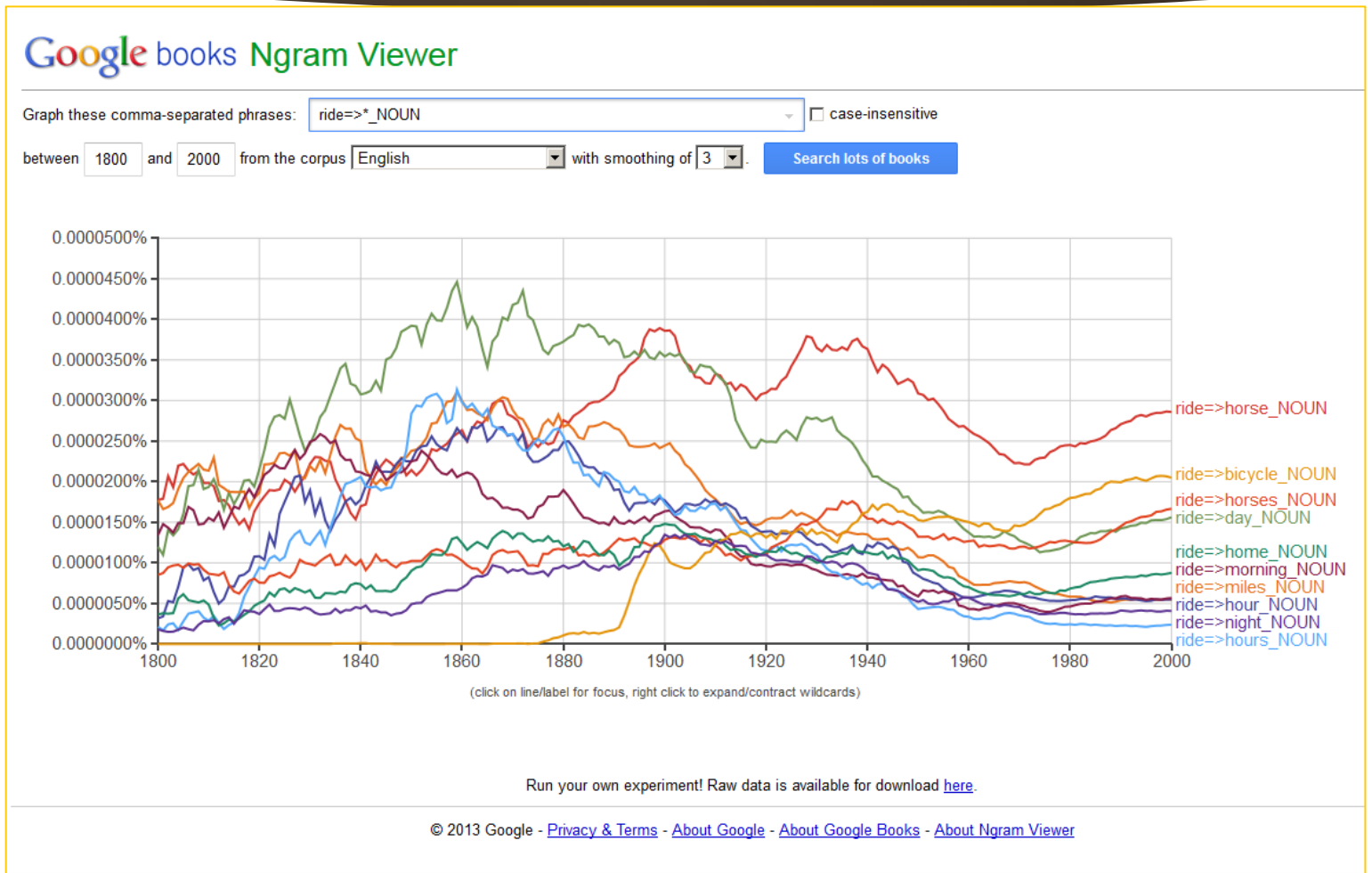
Search in Google Books:

1800 - 1840	1841 - 1931	1932 - 1943	1944 - 1987	1988 - 2000	buy	English
1800 - 1828	1829 - 1931	1932 - 1945	1946 - 1981	1982 - 2000	bought	English
1800 - 1851	1852 - 1936	1937 - 1946	1947 - 1986	1987 - 2000	buying	English
1800 - 1841	1842 - 1926	1927 - 1938	1939 - 1984	1985 - 2000	buys	English

Excluding ngrams with part-of-speech tags from search links.

Run your own experiment! Raw data is available for download here.

<ride=>*_NOUN>


Google books Ngram Viewer
Graph these comma-separated phrases: ride=>*_NOUN
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3 .
Search lots of books
0.0000500%
0.0000450%
0.0000400%
0.0000350%
0.0000300%
0.0000250%
0.0000200%
0.0000150%
0.0000100%
0.0000050%
0.0000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
ride=>horse_NOUN
ride=>bicycle_NOUN
ride=>horses_NOUN
ride=>day_NOUN
ride=>home_NOUN
ride=>morning_NOUN
ride=>miles_NOUN
ride=>hour_NOUN
ride=>night_NOUN
ride=>hours_NOUN
(click on line/label for focus, right click to expand/contract wildcards)
Run your own experiment! Raw data is available for download here.
© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

Ngrams at the Starts and Ends of Sentences

- Sentence Boundary Indicators
  - _START_
  - _END_

<_START_ The USSR,_START_ The US>

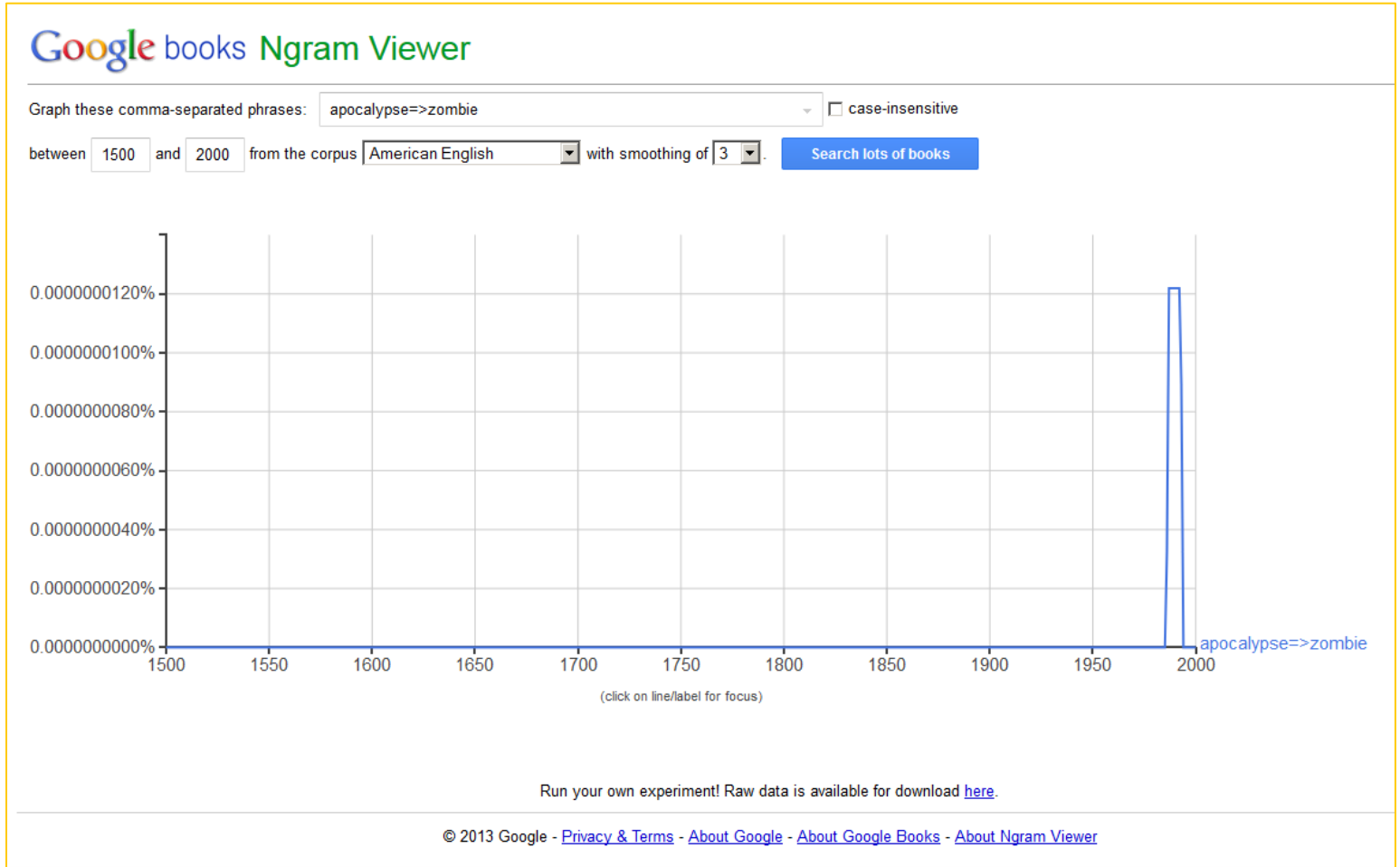
Dependency Relations with => Operator

- main noun => descriptor (the main noun dependent on the descriptor)

<home=>sweet>


Google books Ngram Viewer
Graph these comma-separated phrases: home=>sweet
case-insensitive
between 1700 and 2000 from the corpus English with smoothing of 3 .
Search lots of books
0.0000260%
0.0000240%
0.0000220%
0.0000200%
0.0000180%
0.0000160%
0.0000140%
0.0000120%
0.0000100%
0.0000080%
0.0000060%
0.0000040%
0.0000020%
0.0000000%
1700
1750
1800
1850
1900
1950
2000
home=>sweet
(click on line/label for focus)
Run your own experiment! Raw data is available for download here.
© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

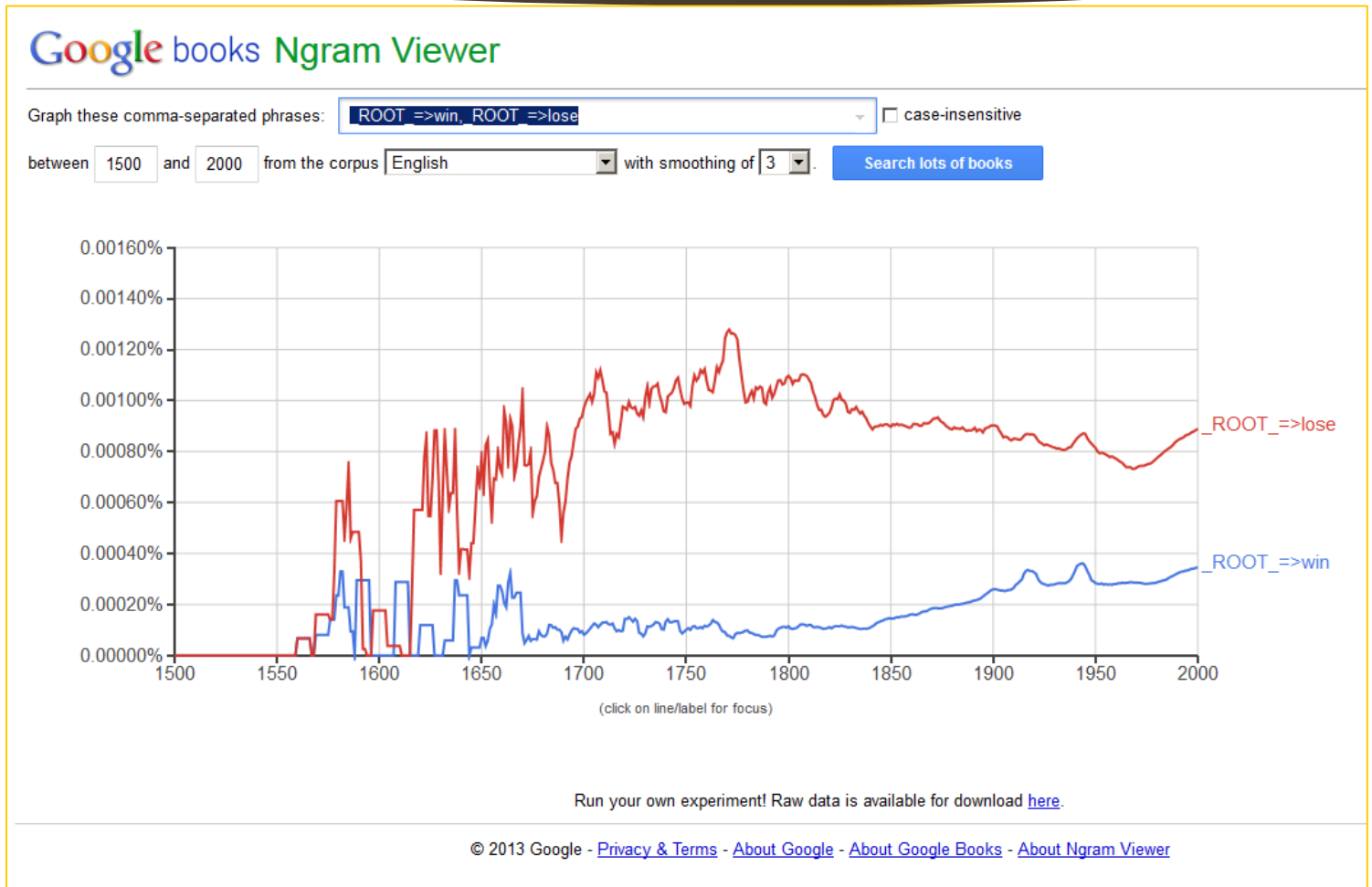
<apocalypse=>zombie>



Root: _ROOT_

- Stands for the root of the parse tree (syntax tree) connected based on the syntax
- Placeholder for "what the main verb of the sentence is modifying" ("Google Books Ngram Viewer")
- Does not stand in for a word or a position in a sentence

<_ROOT_=>win,_ROOT_=>lose>



Ngram Compositions

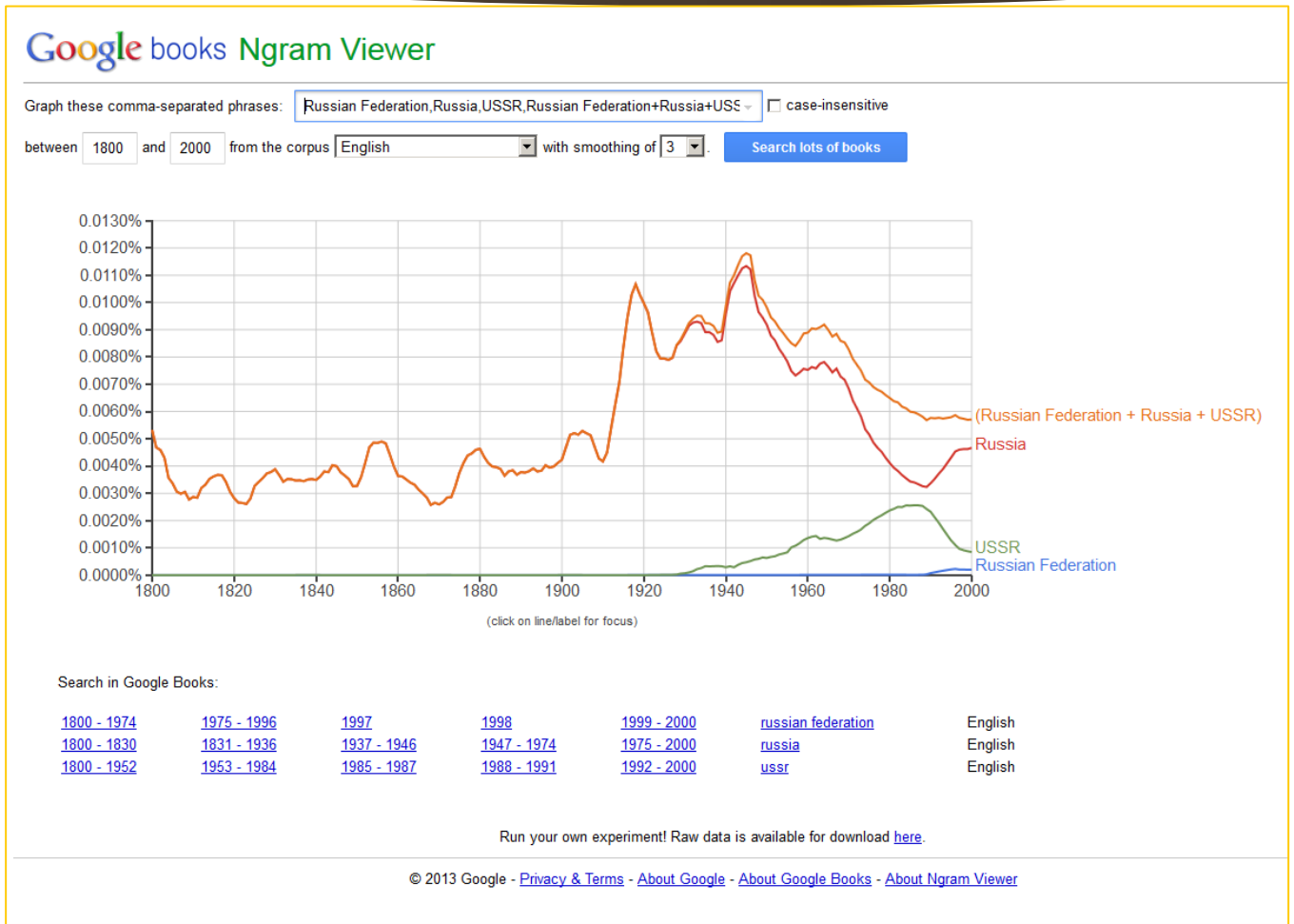
USING OPERATORS

( )

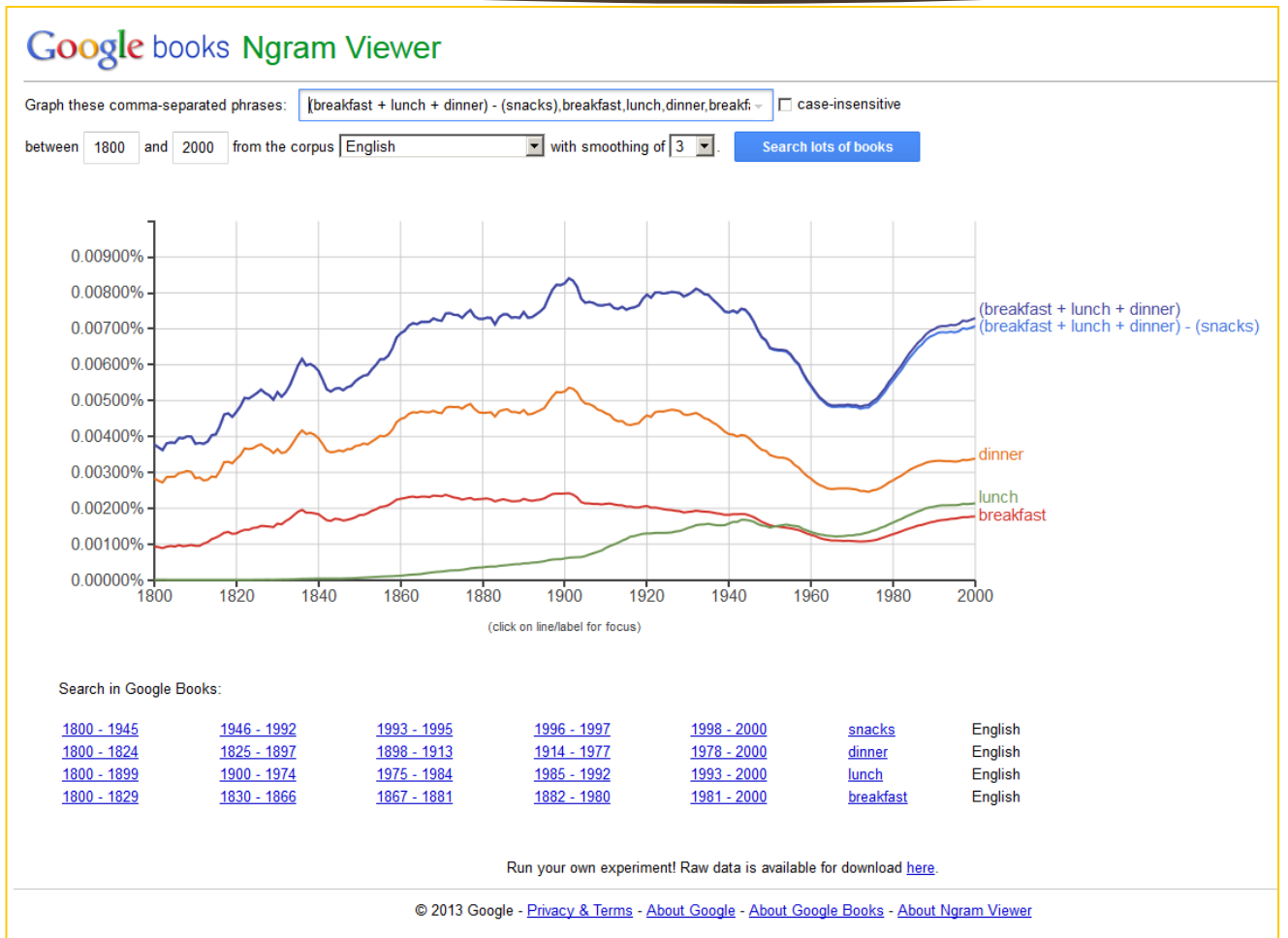
Operators ( )

Operators	Functions
+	sums expressions on the left and the right to combine multiple ngram time series into one line in the linegraph (can combine multiple added sequences)
-	subtracts expression on right from the left (need spaces on either side of the minus sign)
/	divides string on left by expression on right
*	multiplies expression on left by number on right to compare ngrams of different frequencies with entire ngram in ( ) so the asterisk is seen as a multiplication sign and function
:	applies ngram on left to the text corpus on the right to enable comparisons against different corpuses

+ operator <Russian Federation, Russia, USSR, RussianFederation+Russia+USSR>



- operator (+) – (+) <breakfast,lunch,dinner,breakfast+lunch+dinner,(breakfast+lunch+dinner)-(snacks)>



/ operator <traumatic brain injury, TBI, mTBI,(traumatic brain injury / (traumatic brain injury+TBI+mTBI)>

* multiplication operator <HUMINT,GEOINT,MASINT,OSINT,SIGINT,TECHINT,(CYBINT*1000),DNINT,(FINNT*1000)>

- To explore visualizations between texts with widely varying frequencies

The Data Extractions

- HUMINT,GEOINT,MASINT,OSINT,SIGINT,TECHINT,CYBINT,DNINT,FINNT
- HUMINT,GEOINT,MASINT,OSINT,SIGINT,TECHINT,(CYBINT*1000),DNINT,(FINNT*1000) (to no avail since "CYBINT" and "FINNT" do not have sufficient occurrences to count in the Ngram Viewer)

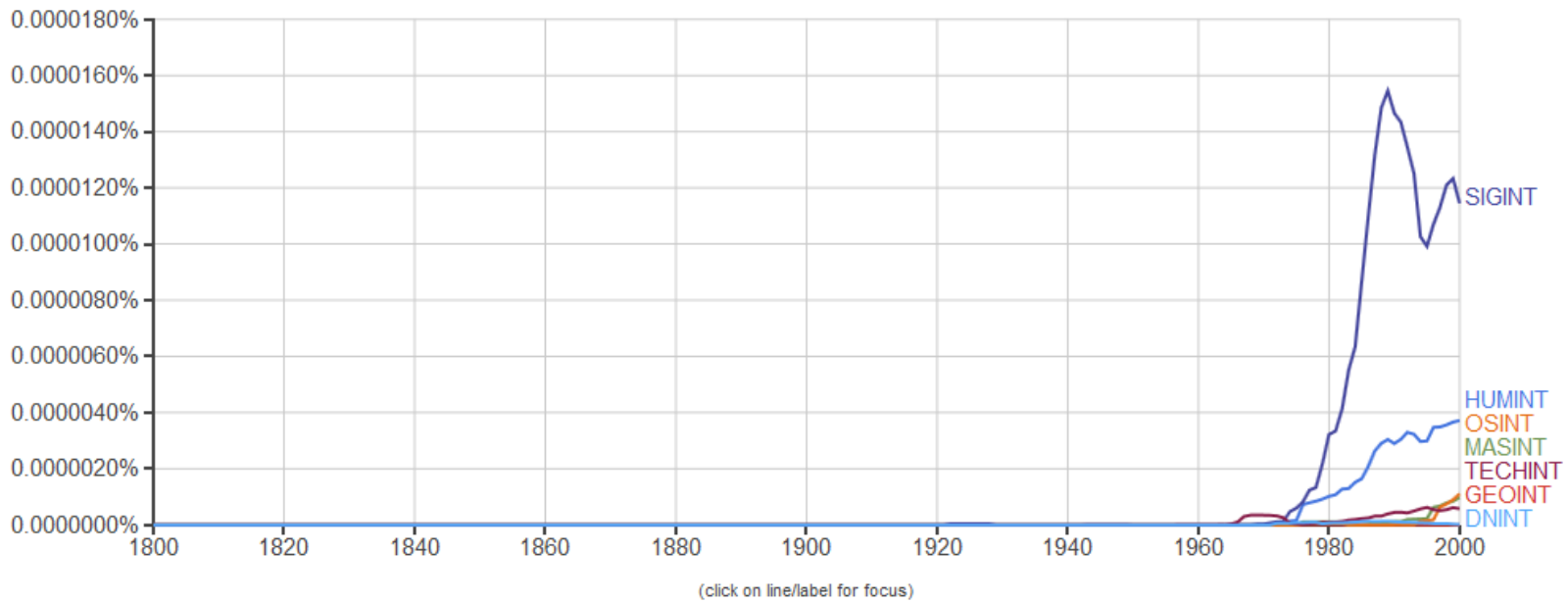
Google books Ngram Viewer

Graph these comma-separated phrases: HUMINT,GEOINT,MASINT,OSINT,SIGINT,TECHINT,(CYBINT*1000),D case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books

Ngrams not found: CYBINT, (CYBINT * 1000), FINNT, (FINNT * 1000)

The characters +, -, *, / require parentheses to be interpreted as a composition.



(click on line/label for focus)

Search in Google Books:

1800 - 1981	1982 - 1989	1990	1991 - 1998	1999 - 2000	humint	English
1800 - 1978	1979 - 1985	1986	1987 - 1990	1991	geoint	English
1800 - 1979	1980 - 1997	1998	1999	2000	masint	English
1800 - 1996	1997	1998 - 1999	2000	osint	English	
1800 - 1983	1984 - 1989	1990	1991 - 1999	2000	sigint	English
1800 - 1969	1970	1971 - 1990	1991 - 1997	1998 - 2000	techint	English
1800 - 1975	1976 - 1987	1988	1989 - 1995	1996 - 2000	dnint	English

: operator <Beijing:eng_gb_2012,Beijing:chi_sim_2012, Beijing:eng_us_2012>

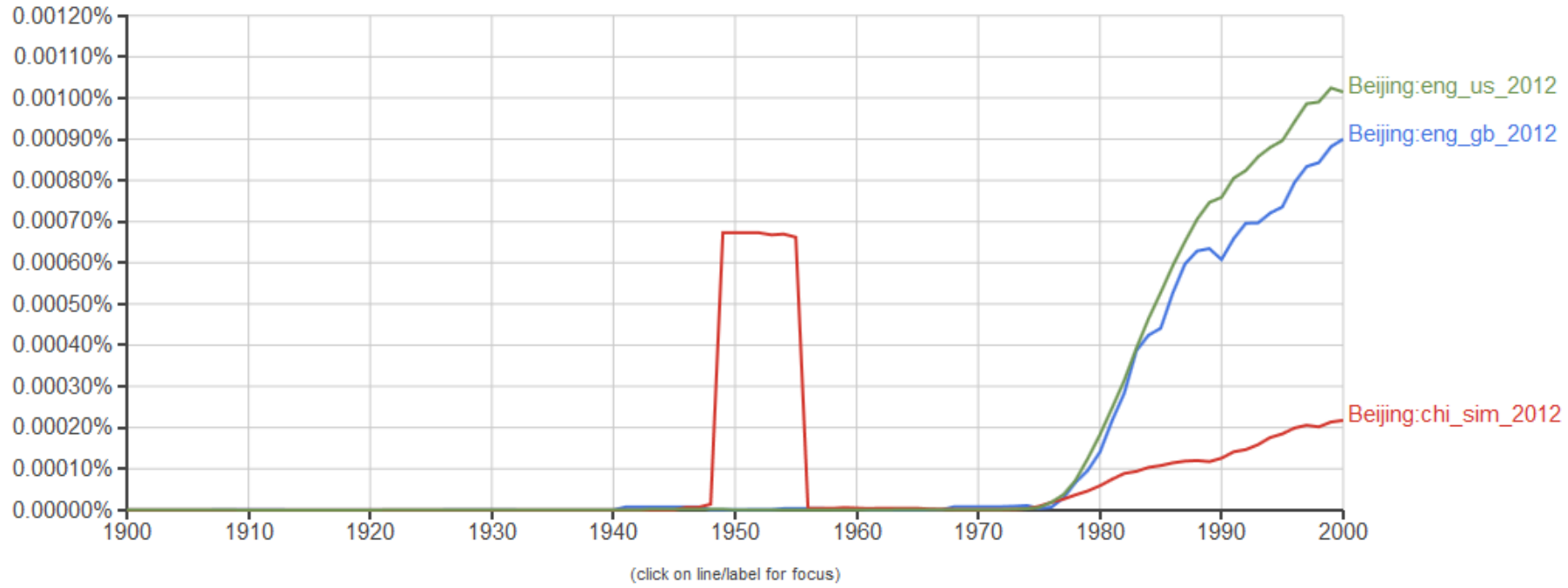
- Beijing:eng_gb_2012,Beijing:chi_sim_2012,Beijing:eng_us_2012

- Some definitions of the pointed-to datasets
  - eng_gb_2012: "Books predominantly in the English language that were published in Great Britain"
  - chi_sim_2012: "Books predominantly in simplified Chinese script"
  - eng_us_2012: "Books predominantly in the English language that were published in the United States"

Google books Ngram Viewer

Graph these comma-separated phrases: Beijing:eng_gb_2012,Beijing:chi_sim_2012,Beijing:eng_us_2012 case-insensitive

between 1900 and 2000 from the corpus English with smoothing of 3. Search lots of books



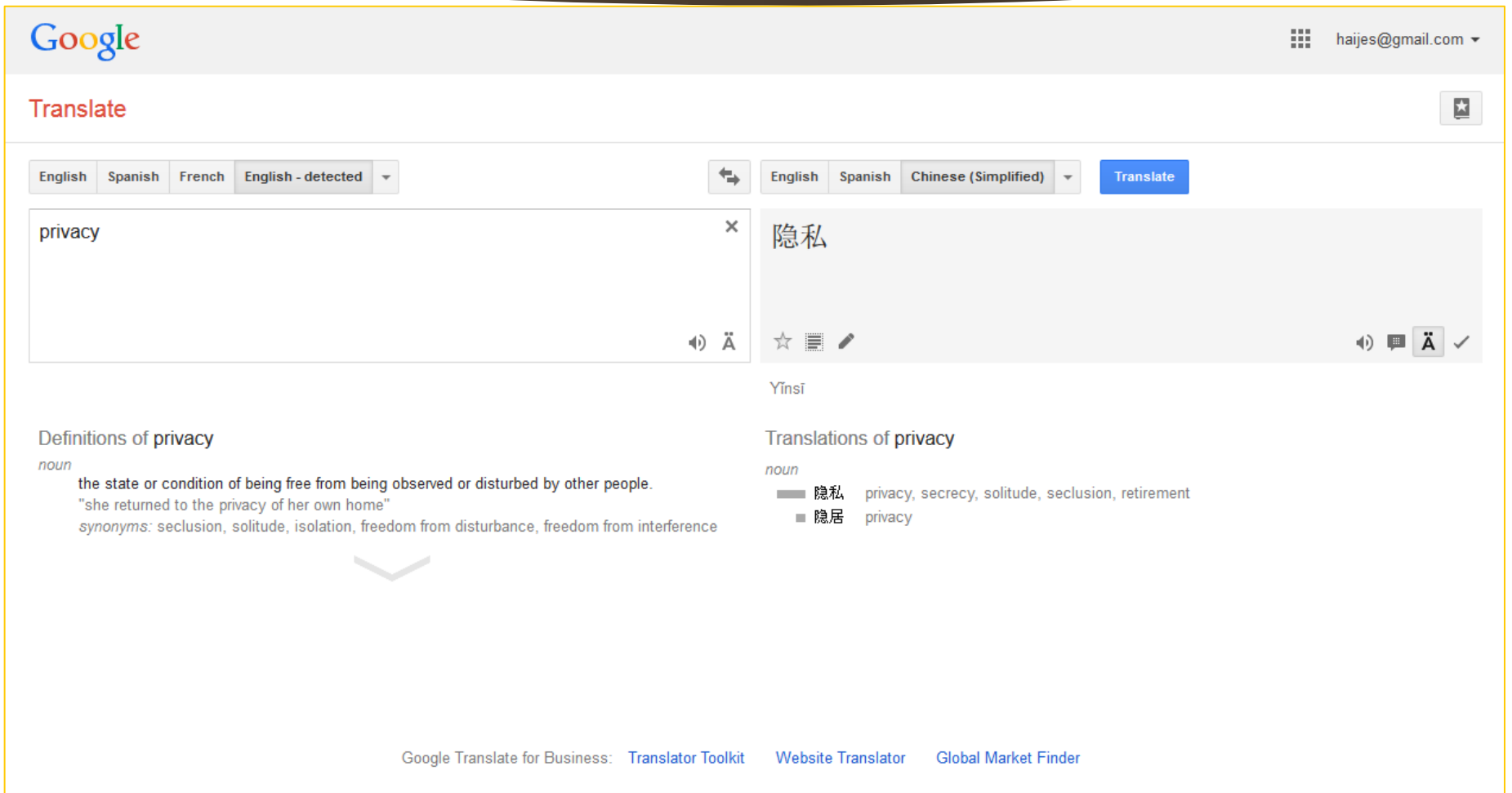
(click on line/label for focus)

Search in Google Books:

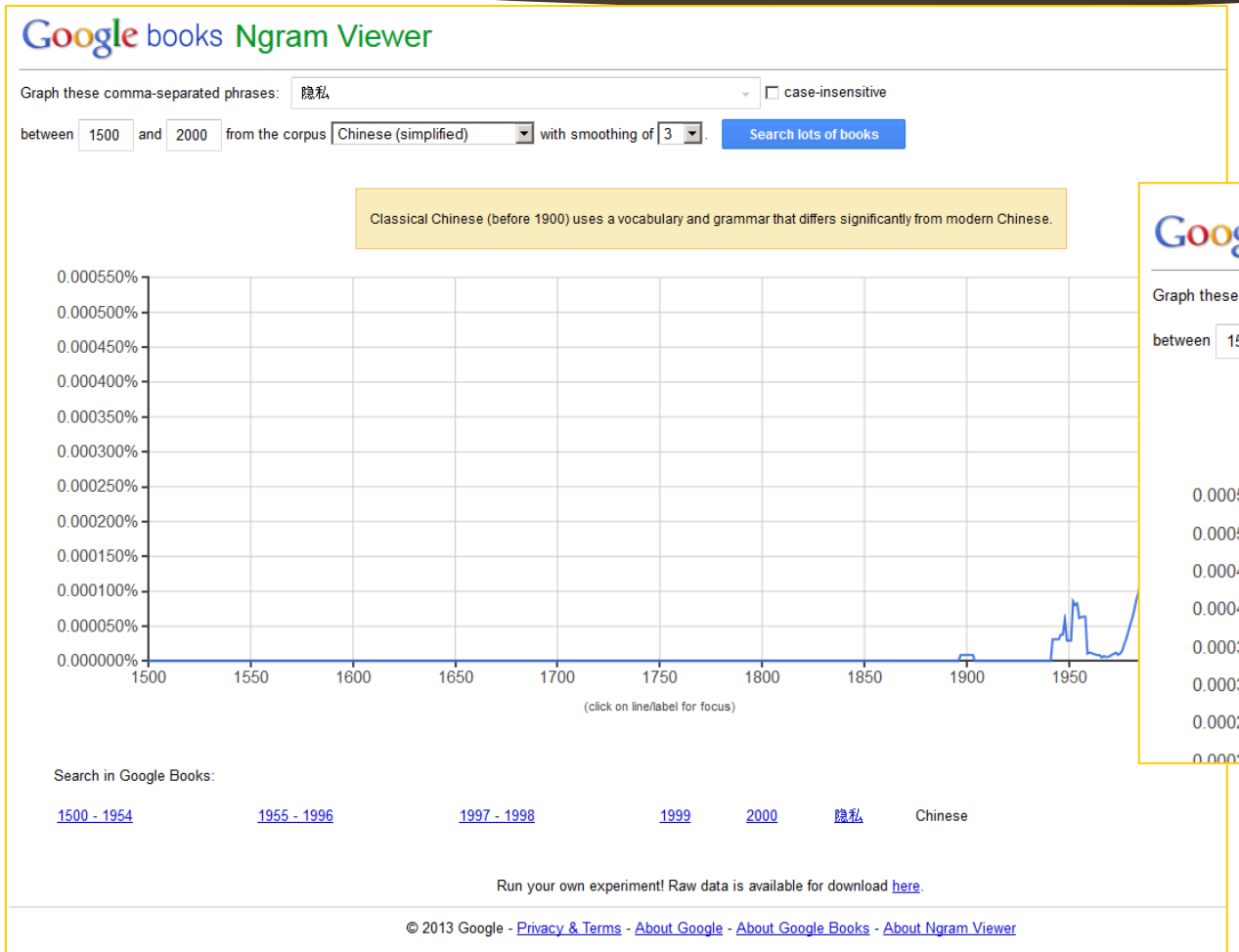
1900 - 1983	1984 - 1995	1996 - 1997	1998	1999 - 2000	beijing	English
1900 - 1951	1952	1953 - 1992	1993 - 2000	2001 - 2000	beijing	Chinese

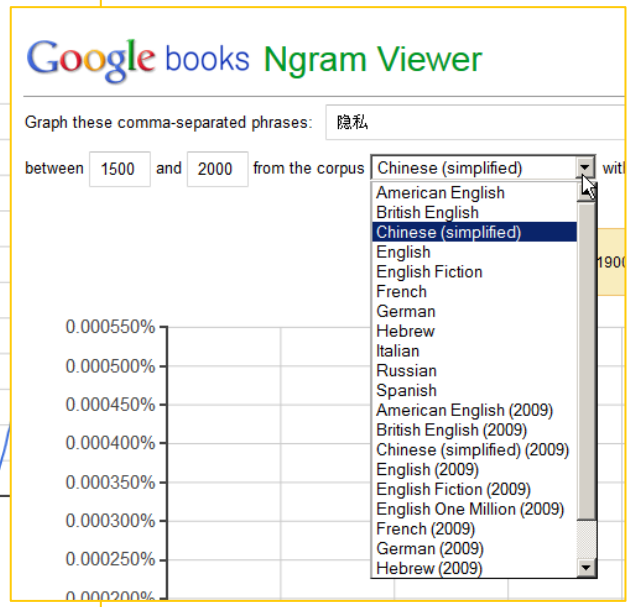
Run your own experiment! Raw data is available for download here.

Two Steps to Privacy: 隐私

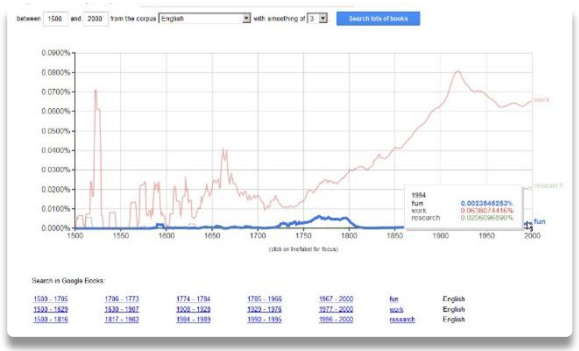


Popularization of Term with Broader Interactions Globally <隐私>



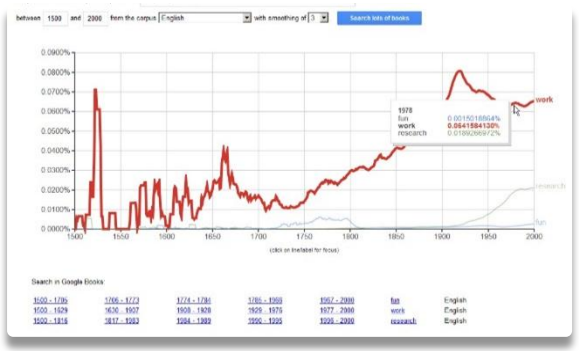


Any Ideas for Using the Ngram Viewer for your Fun, Work, and Research?



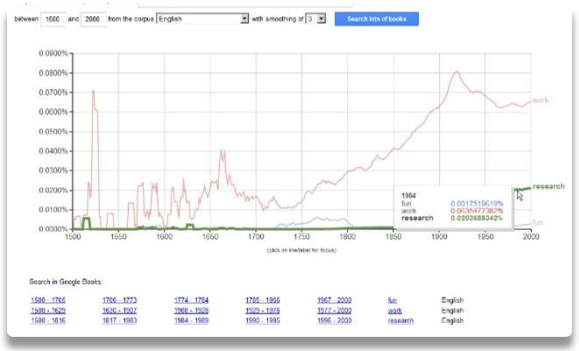
Fun

- Stories to tell friends



Work

- Insights that may affect knowledge and decision-making



Research

- Citable information from the world's collective knowledge conveyed through books…

Ngram Viewer Applications for Visual Wordplay and Wit

A ONE-SCREEN TEXT-BASED VISUAL TO ACCENTUATE A WEBSITE, PRESENTATION, OR PUBLICATION?

<d'oh>

Google books Ngram Viewer

Graph these comma-separated phrases: d'oh case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books

Replaced **d'oh** with **d' oh** to match how we processed the books.

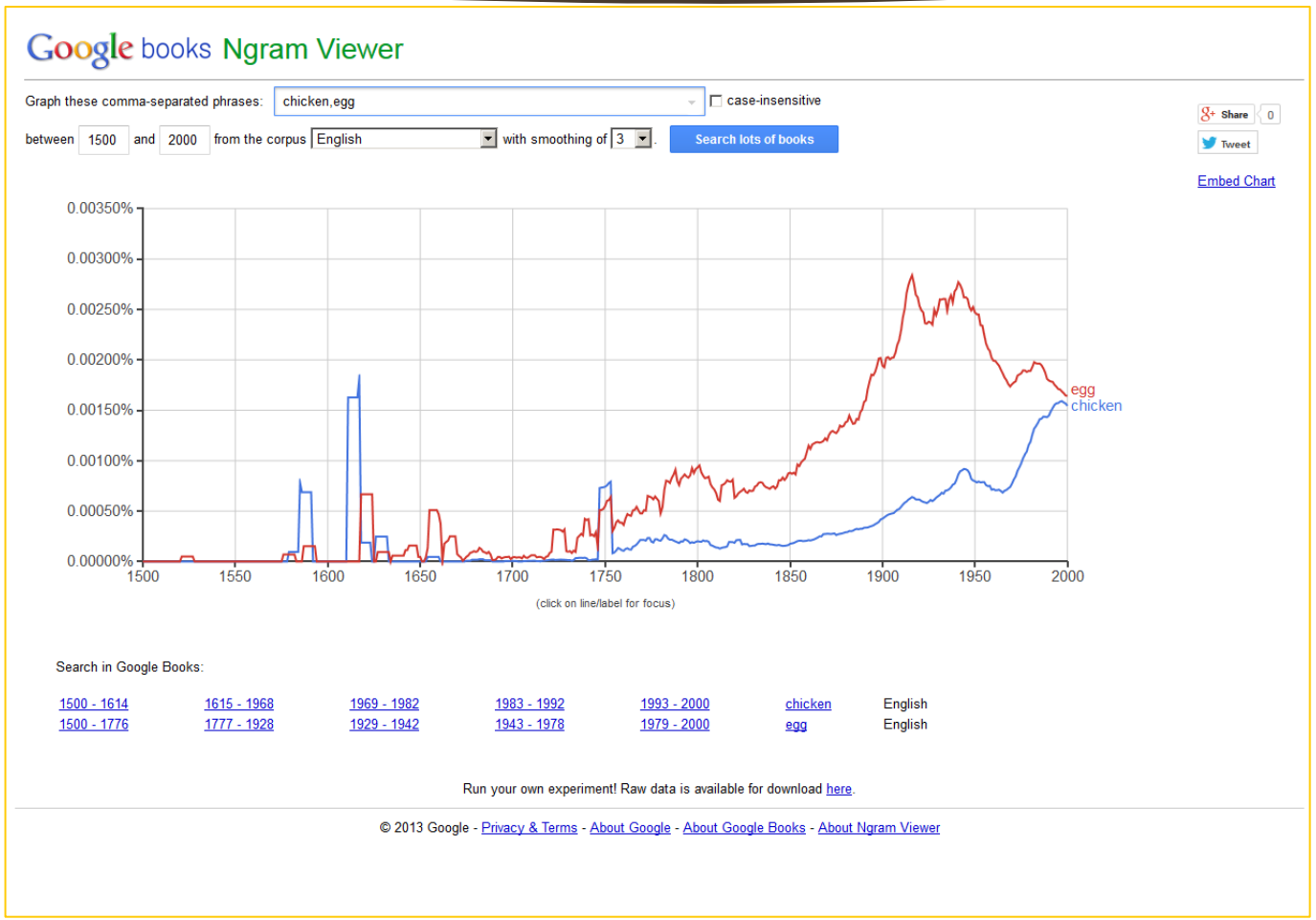
(click on line/label for focus)

Search in Google Books:

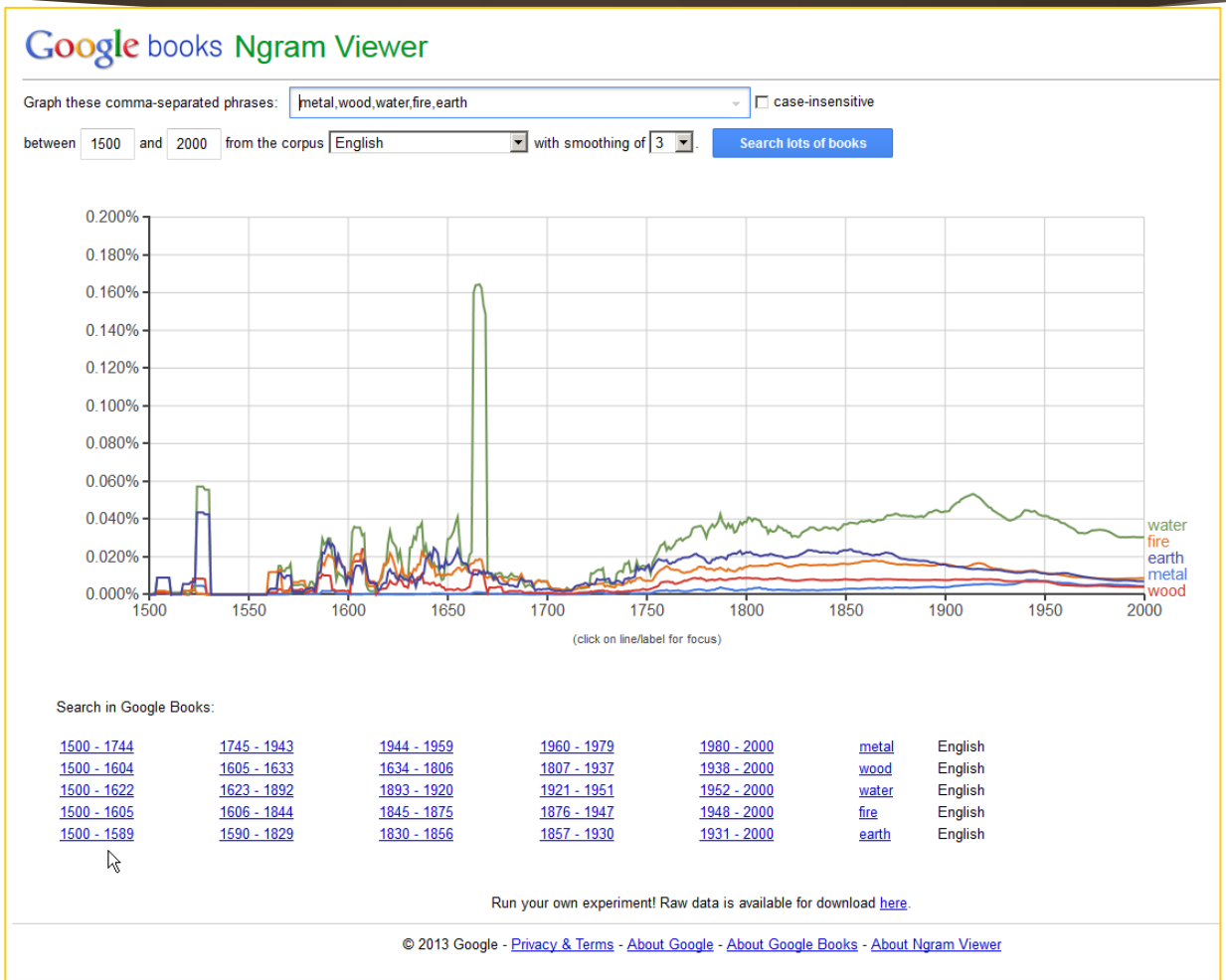
1800 - 1811	1812 - 1905	1906 - 1912	1913 - 1998	1999 - 2000	d'oh	English
1800 - 1811	1812 - 1905	1906 - 1912	1913 - 1998	1999 - 2000	d' oh	English

Run your own experiment! Raw data is available for download here.

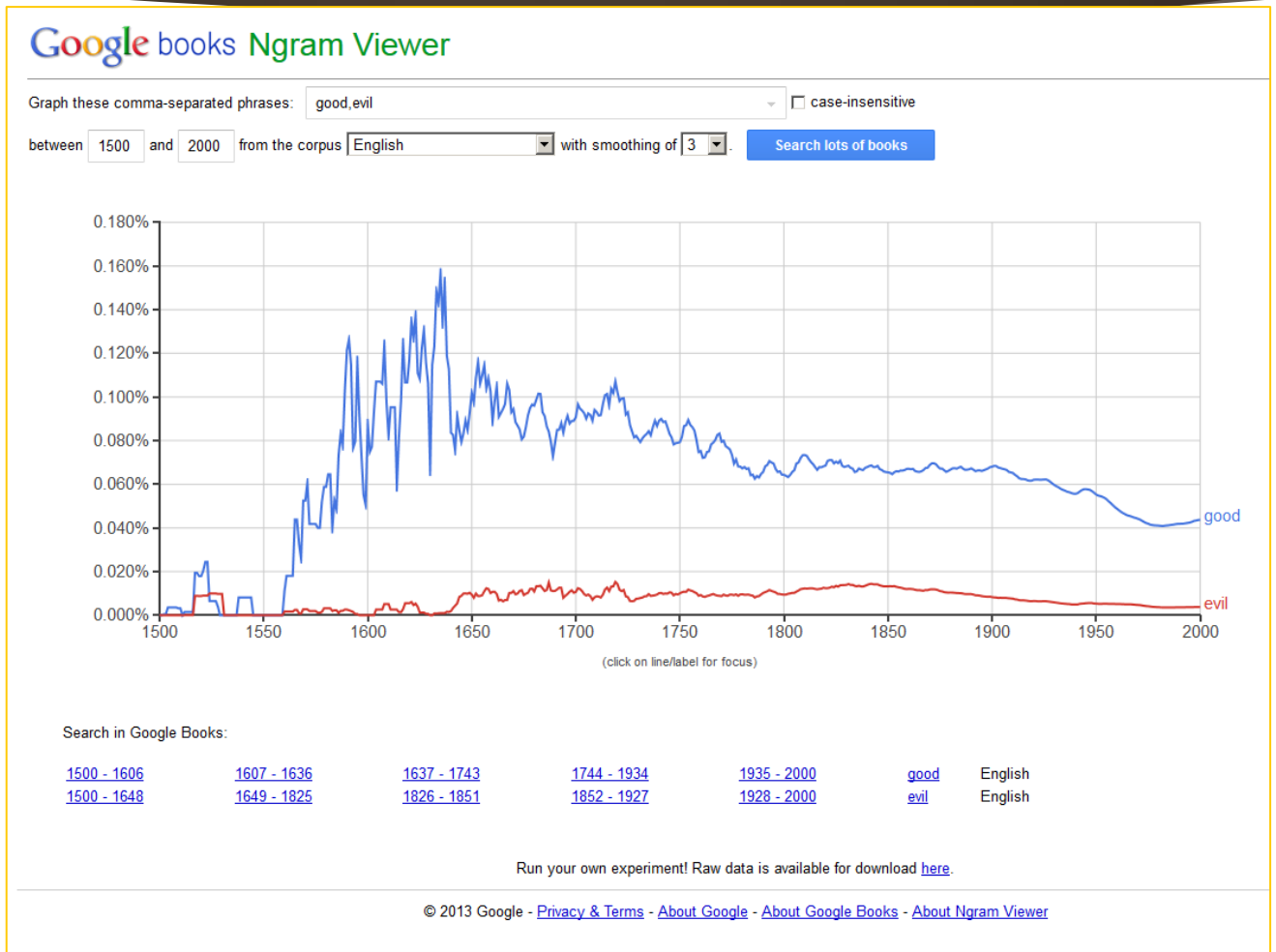
Chicken or the Egg? <chicken,egg>



Five Elements <metal,wood,water,fire,earth>



Good and Evil
<good,evil>



<Buddhism, Christianity, Hinduism, Islam,Judaism, Mormonism, Sikhism>

<Oriental, Asian>

Google books Ngram Viewer

Graph these comma-separated phrases: Oriental,Asian case-insensitive

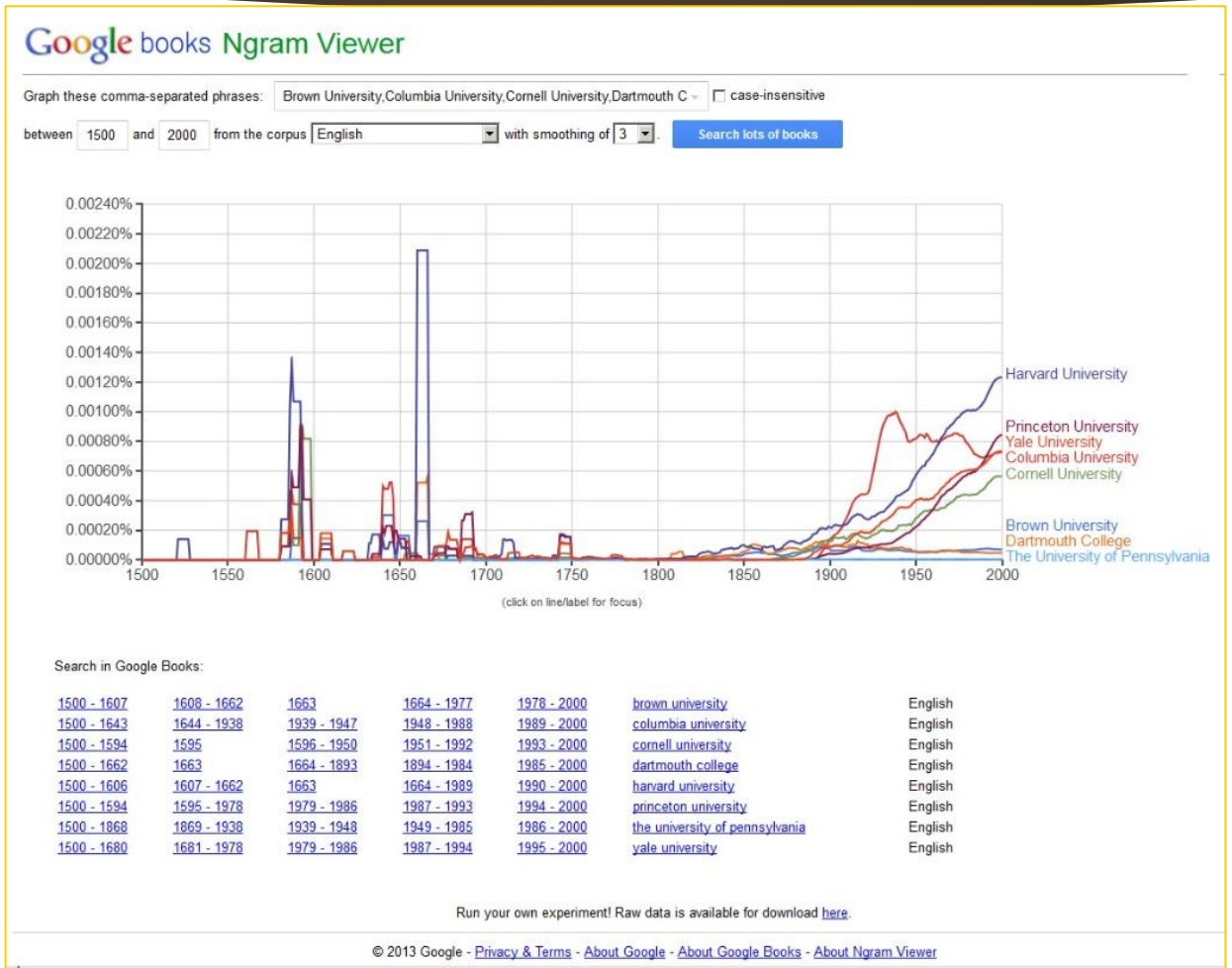
between 1500 and 2000 from the corpus English with smoothing of 3 . Search lots of books

Search in Google Books:

1500 - 1698	1699 - 1920	1921 - 1941	1942 - 1967	1968 - 2000	oriental	English
1500 - 1738	1739 - 1988	1989 - 1992	1993 - 1995	1996 - 2000	asian	English

Run your own experiment! Raw data is available for download here.

Ivy League Institutions <Brown University, Columbia University, Cornell University, Dartmouth College, Harvard University, Princeton University, the University of Pennsylvania, Yale University>



Caveat

- The ease of accessing and understanding the visualization may mean a potential misunderstanding of the underlying information… This is partly a product of the cognitive bias known as the "availability heuristic," with more easeful and faster ideas coming to mind accepted as truth.
- Visualizations like this are highly overly simplified as compared to the underlying realities.
- Researchers need to make sure to head off potential misunderstandings with such Ngram Viewer linegraph visualizations when using these as "accents" or as "invitations" to people to learn more.

Ngram Viewer Uses in Research

EARLY THOUGHTS

Some Possible Research Applications of the Ngram Viewer

- Variations of the following:
  - Competition between languages and phrases (their origins and trajectories / trends over time, word and phrase gists over time, multilingual queries, and others)
  - Cultural understandings and cross-cultural insights; popular sentiment and understandings
  - Analysis of research capabilities and understandings (historically and through the present)
  - Population readiness for accepting particular ideas through big data text corpus analysis
  - Literary terms of art and their uses over time

Some Possible Research Applications of the Ngram Viewer (cont.)

- Effects of historical events (governance, social phenomena, wars, health issues, and others) on language
- Biographical insights on historical figures (particularly comparative insights)
- Research lead creation; research source identification
- and many others

Google books Ngram Viewer

Graph these comma-separated phrases: meme ☐ case-insensitive

between 1500 and 2000 from the corpus American English with smoothing of 3 . Search lots of books

Search in Google Books:

1500 - 1578 1579 1580 - 1776 1777 - 1939 1940 - 2000 meme English

Run your own experiment! Raw data is available for download here.

Qualifiers and Clarifications

- Words in books as a lagging (vs. leading) indicator
- Changing authorship and access to literacy and publication over time (with the changing roles of books from years of formalism to much less formalism in the present day)
- Word frequency counts as one information stream among many
  - Still a critical role for close readings of select publications
- Ngram Viewer counts are much more effective and informative when used with complementary streams of information and in-depth analysis

Particular Researcher Requirements

General Understandings

Language literacy, optimal multilingual literacy

Digital (and computational) literacy

Understandings of history

Understandings of the changing roles of authors and books

Understandings of big data and big data analyses

Domain + Computational Understandings

Deep knowledge of particular domain fields and related fields

Understandings of language uses and publications in-the-field

Computational set thinking

Uses of the Ngram Viewer Tool

Openness to discovery learning

Knowing what to query and how (particularly with creative query setups, year adjustment parameters, links to documents)

Knowing what may be asserted and what may not be asserted (ability to qualify assertions)

Knowing when to conduct complementary and follow-on research (including close readings)

Initial Ngram Viewer Tips

- Start simple. Once the basic extractions are acquired, try the more complex ones using tagging and combinatorial approaches. Broaden out to foreign languages.
- Reload the Ngram Viewer if a "flatline" is attained because "under heavy load, the Ngram Viewer will sometimes return a flatline".
- Text corpuses accessed by the Ngram Viewer are always changing, and more data is added all the time. It may help to capture a sequence of data extractions to what changes there may be.

Tips on Research Approaches

- Shape a data query both for need-to-know and to the limits of the massive dataset.
- Err on the side of making a number of various runs for a data query. Keep good records of the data extractions.
- Take time to actually analyze the results. Sometimes, because the extractions occur in milliseconds, just making a cursory look at the linegraph seems sufficient…but much more can be learned by interacting with the visual. (One can explore years, resources, and other aspects, for example.)
- Keep a research journal of observations and findings. Note your learning about the tool as well.

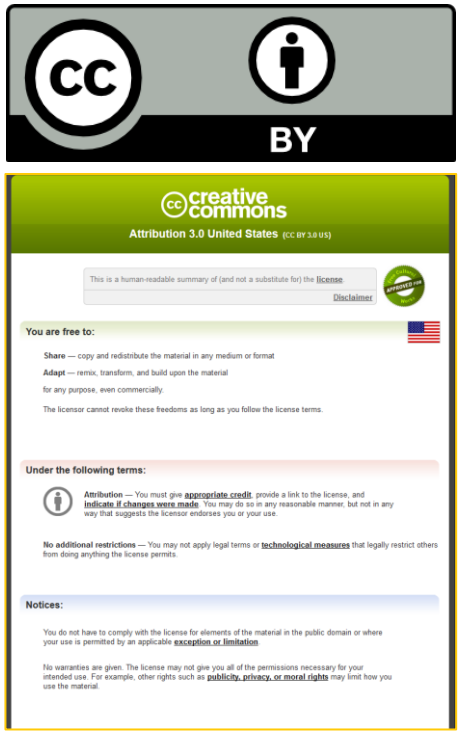
Tips on Research Approaches (cont.)

- Spend some time to discover the tool by making various runs purely for discovery learning. Structure some of these explorations with thought experiments.
- Branch out beyond the Ngram Viewer by analyzing the extracted datasets in other tools. One freeware and open-source one is the Ngram Statistics Package.
- Also, use other datasets (such as from social media platform-extracted big data corpuses) for analysis. One such publicly available set is the Rovereto Twitter N-Gram Corpus.

Tips on Research Approaches (cont.)

- There is a broad and wide literature on the machine analysis of human language: natural language processing, stylometry, computational linguistics, sentiment analysis, personality analysis, speech recognition, and others. There are automated text summaries (with efforts towards accuracy and "grammaticality"). There are language models used for speech recognition and machine translation between languages. A core unit underlying these approaches are n-grams. It may help to delve more deeply for certain types of research to more fully contextualize research-based approaches.

Creative Commons Release

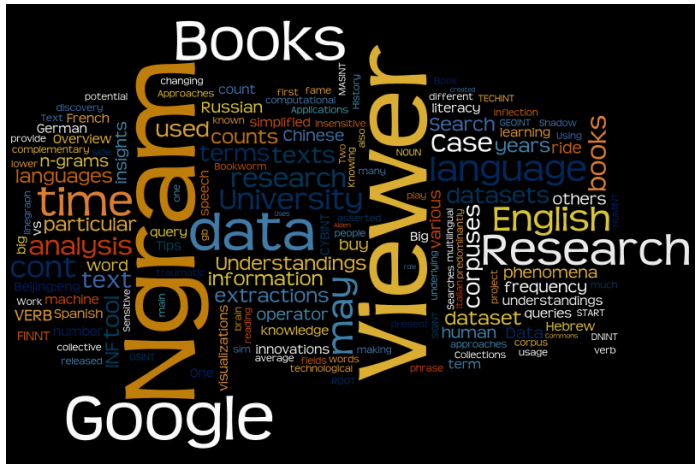


- Currently, datasets and graphs are released through a Creative Commons Attribution 3.0 Unported License.
- Graphs may be used "freely…for any purpose" albeit acknowledgment of Google Books Ngram Viewer and link to http://books.google.com/ngrams is desirable.

References

- Aiden, E. & Michel, J.-B. (2013). *Uncharted: Big Data as a Lens on Human Culture*. New York: Riverhead Books.
- Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution that will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt Publishing Company.

Conclusion and Contact



- Dr. Shalin Hai-Jew
  - Instructional Designer
- Information Technology Assistance Center
- Kansas State University
- 212 Hale Library
- 785-532-5262
- shalin@k-state.edu