Ngrams, Word Frequencies, and the Neural Model

Philip Garnett
York Cross-disciplinary Centre for Systems Analysis and School of Management
philip.garnett@york.ac.uk
@prgarnett (twitter)

Or… Word Frequency Analysis for Measuring Change

Collaborative Effort

- This work sort of drifts along at its own pace as a cross-disciplinary effort between a group of people that drop in and out of papers.
- We are mostly scientists, and not linguists.
- This is interesting in itself as how some types of research get done.

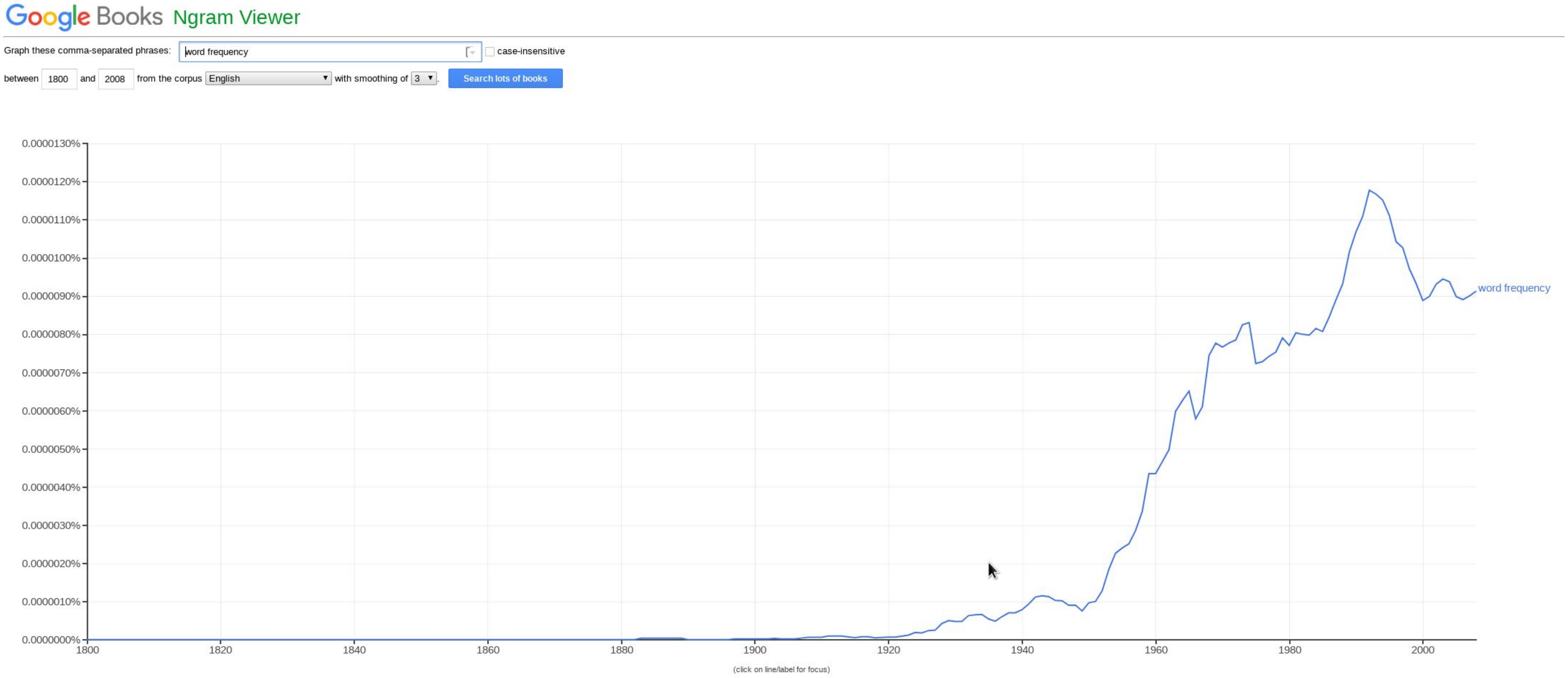
- Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. “The Expression of Emotions in 20th Century Books.” PloS One 8 (3): e59030.
- Bentley, R. A., Eleanor J. Maddison, P. H. Ranner, John Bissell, Camila C. S. Caiado, Pojanath Bhatanacharoen, Timothy Clark, et al. 2014. “Social Tipping Points and Earth Systems Dynamics.” Frontiers of Environmental Science & Engineering in China 2: 35.
- Bentley, R. Alexander, Philip Garnett, Michael J. O’Brien, and William A. Brock. 2012. “Word Diffusion and Climate Science.” PloS One 7 (11): e47966.
- Clark, Timothy, Mike Wright, Zilia Iskoujina, and Philip Garnett. 2014. “JMS at 50: Trends over Time.” Journal of Management Studies 51 (1): 19–37.
- Ruck, Damian, R. Alexander Bentley, Alberto Acerbi, Philip Garnett, and Daniel J. Hruschka. 2017. “ROLE OF NEUTRAL EVOLUTION IN WORD TURNOVER DURING CENTURIES OF ENGLISH WORD POPULARITY.” Advances in Complex Systems 20 (06n07): 1750012.
- Skrebyte, Agne, Philip Garnett, and Jeremy R. Kendal. 2016. “Temporal Relationships Between Individualism–Collectivism and the Economy in Soviet Russia: A Word Frequency Analysis Using the Google Ngram Corpus.” Journal of Cross-Cultural Psychology 47 (9): 1217–35.

Ngrams Data

Google books

- As Google sometimes does it embarked on a ambitious plan to scan all the World’s books.
    - Would have been an amazing resource.
    - Instantly got bogged down in legal problems.
    - Did scan ~25 million books largely from University Libraries.
    - Basis of the Google Books web tool.
    - Also produced Ngrams...

Ngrams



Google Books Ngram Viewer

Graph these comma-separated phrases: word frequency □ case-insensitive

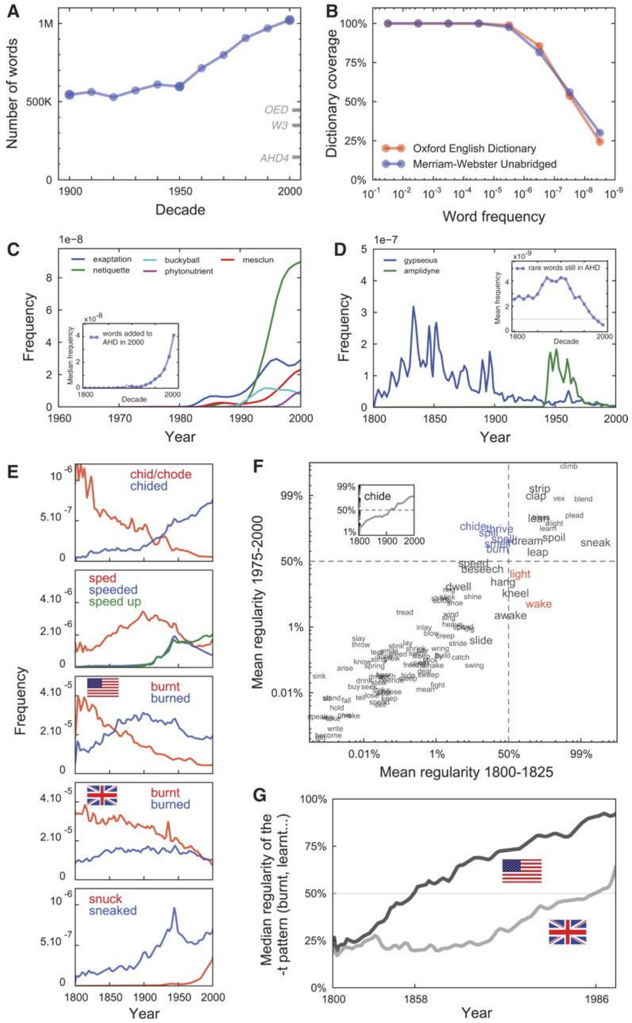
between 1800 and 2008 from the corpus English with smoothing of 3.

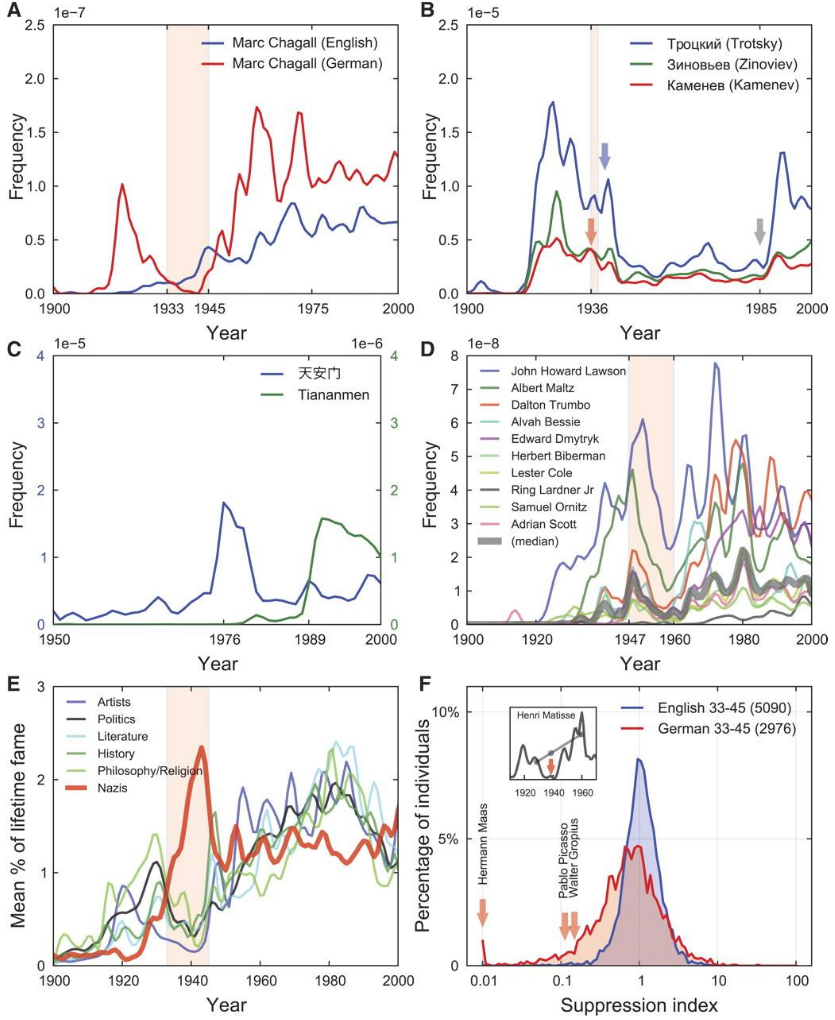
Culturomics

“Culturomics is the application of high-throughput data collection and analysis to the study of human culture. Books are a beginning, but we must also incorporate newspapers (29), manuscripts (30), maps (31), artwork (32), and a myriad of other human creations (33, 34). Of course, many voices—already lost to time—lie forever beyond our reach.

Culturomic results are a new type of evidence in the humanities. As with fossils of ancient creatures, the challenge of culturomics lies in the interpretation of this evidence. Considerations of space restrict us to the briefest of surveys: a handful of trajectories and our initial interpretations. Many more fossils (Fig. 5 and fig. S13), with shapes no less intriguing, beckon…”

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, et al. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books.” Science 331 (6014): 176–82.

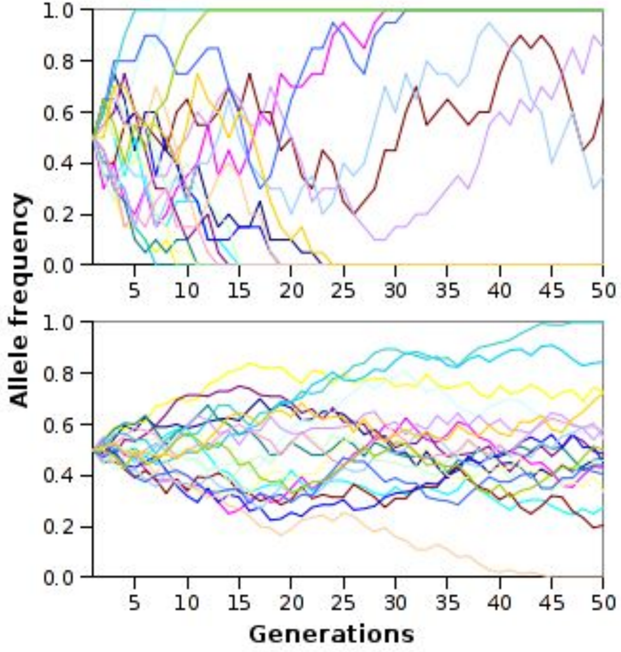




Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." Science 331 (6014): 176–82.

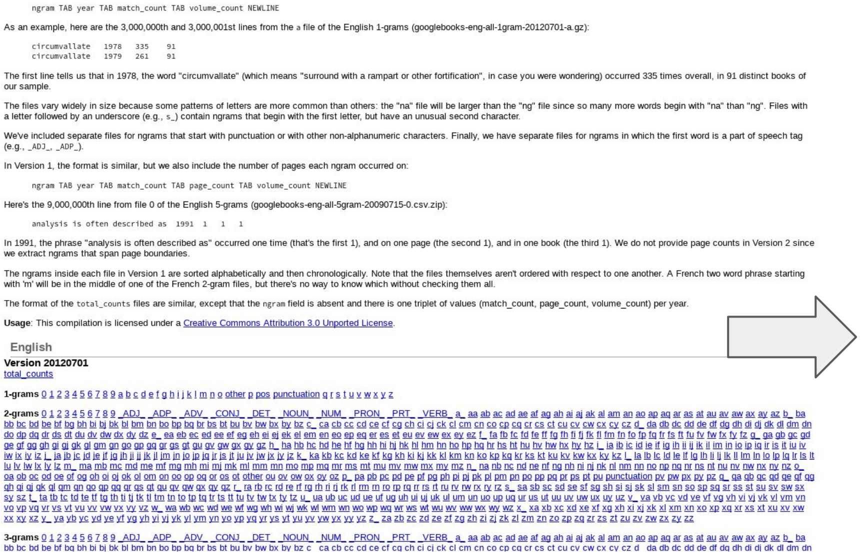
Culturomics and Genomics

- Culturomics sounds a bit like genomics… and they share a few traits.
- I was a Geneticist that got interested in people…
- One of things that Geneticists spend a lot of time worrying about is selection vs drift.
- Both are ways things adapt - hard to tell apart.



You Can Build You Own Ngram Tools

We downloaded it and then shoved it in a MySQL database so we could read it out using computers easily.



```
ngram TAB year TAB match_count TAB volume_count NEWLINE
```

As an example, here are the 3,000,000th and 3,000,001st lines from the a file of the English 1-grams (googlebooks-eng-all-1gram-20120701-a.gz):

```
circumvallate   1978   335    91
circumvallate   1979   261    91
```

The first line tells us that in 1978, the word "circumvallate" (which means "surround with a rampart or other fortification", in case you were wondering) occurred 335 times overall, in 91 distinct books of our sample.

The files vary widely in size because some patterns of letters are more common than others: the "na" file will be larger than the "ng" file since so many more words begin with "na" than "ng". Files with a letter followed by an underscore (e.g., s_) contain ngrams that begin with the first letter, but have an unusual second character.

We've included separate files for ngrams that start with punctuation or with other non-alphanumeric characters. Finally, we have separate files for ngrams in which the first word is a part of speech tag (e.g., _ADJ_, _ADP_).

In Version 1, the format is similar, but we also include the number of pages each ngram occurred on:

```
ngram TAB year TAB match_count TAB page_count TAB volume_count NEWLINE
```

Here's the 9,000,000th line from file 0 of the English 5-grams (googlebooks-eng-all-5gram-20090715-0.csv.zip):

```
analysis is often described as  1991  1   1   1
```

In 1991, the phrase "analysis is often described as" occurred one time (that's the first 1), and on one page (the second 1), and in one book (the third 1). We do not provide page counts in Version 2 since we extract ngrams that span page boundaries.

The ngrams inside each file in Version 1 are sorted alphabetically and then chronologically. Note that the files themselves aren't ordered with respect to one another. A French two word phrase starting with 'm' will be in the middle of one of the French 2-gram files, but there's no way to know which without checking them all.

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.

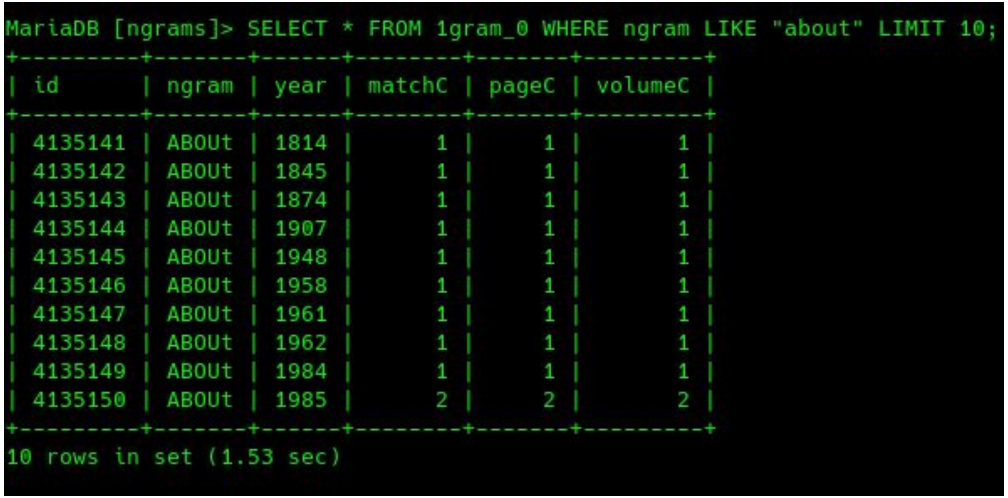
English

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



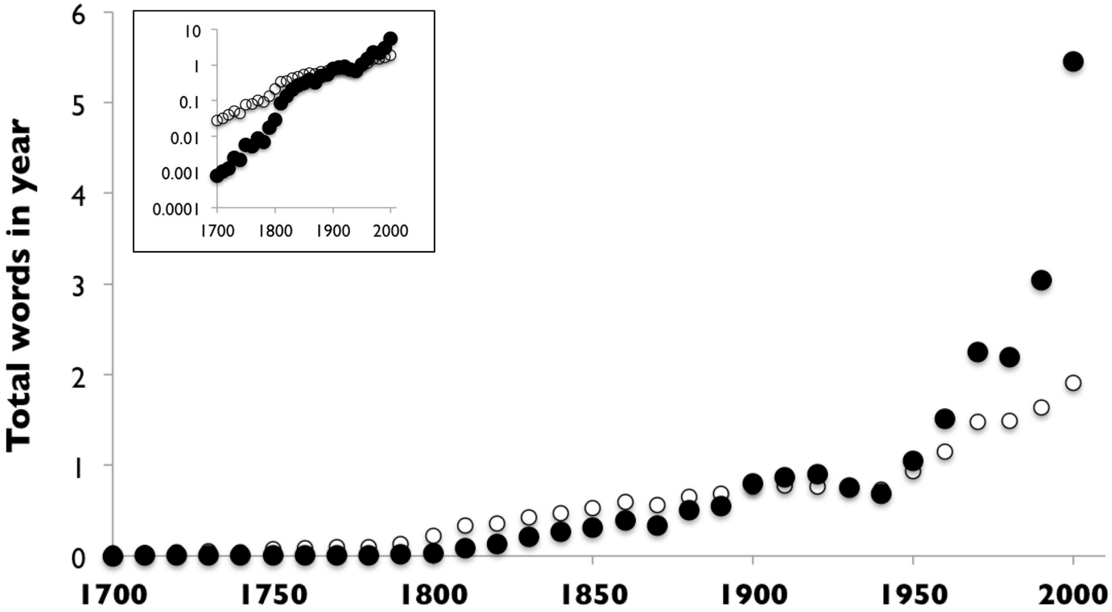
```
MariaDB [ngrams]> SELECT * FROM 1gram_0 WHERE ngram LIKE "about" LIMIT 10;
+---------+-------+------+--------+-------+---------+
| id      | ngram | year | matchC | pageC | volumeC |
+---------+-------+------+--------+-------+---------+
| 4135141 | ABOUt | 1814 |      1 |     1 |       1 |
| 4135142 | ABOUt | 1845 |      1 |     1 |       1 |
| 4135143 | ABOUt | 1874 |      1 |     1 |       1 |
| 4135144 | ABOUt | 1907 |      1 |     1 |       1 |
| 4135145 | ABOUt | 1948 |      1 |     1 |       1 |
| 4135146 | ABOUt | 1958 |      1 |     1 |       1 |
| 4135147 | ABOUt | 1961 |      1 |     1 |       1 |
| 4135148 | ABOUt | 1962 |      1 |     1 |       1 |
| 4135149 | ABOUt | 1984 |      1 |     1 |       1 |
| 4135150 | ABOUt | 1985 |      2 |     2 |       2 |
+---------+-------+------+--------+-------+---------+
10 rows in set (1.53 sec)
```

Ngram Data Set

Runs from ~1700 to 2009.

Bit rubbish from about 1700-1800 (low frequencies & errors)

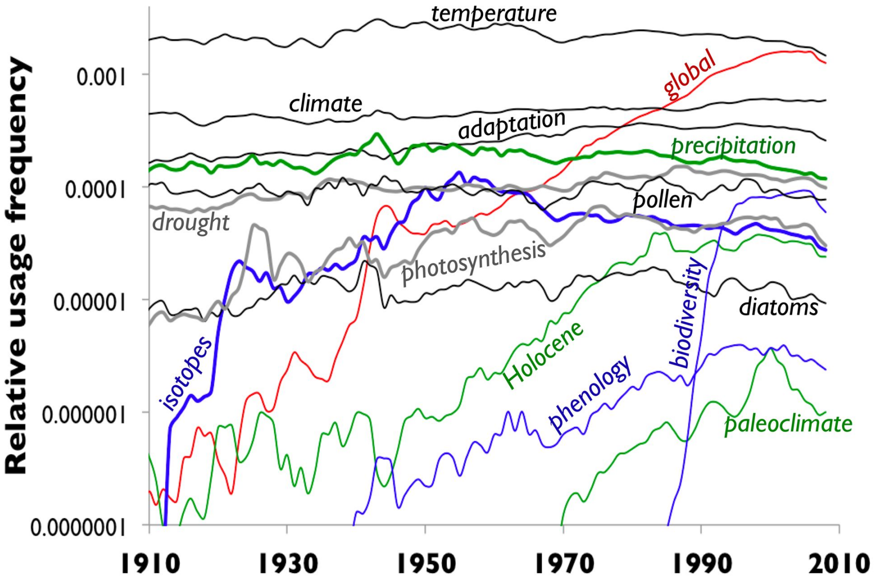
You have to be careful with change of spelling…



Word Frequency Analysis

We could have a play with that!

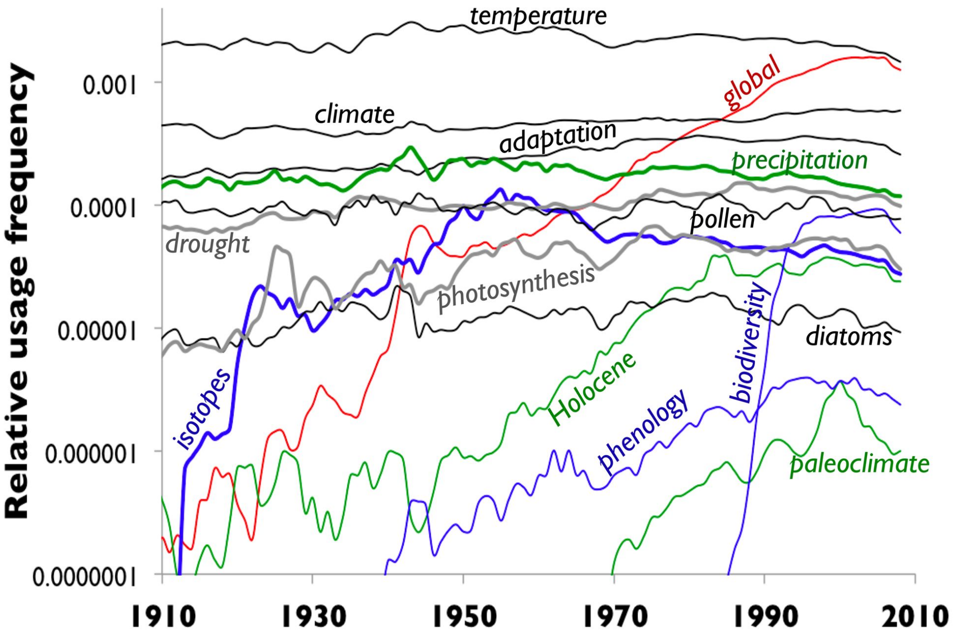
- At the time a bunch of us were working on Tipping Points and interested in climate change.
- Tipping point - rapid shift from one system state to another.
- So we got interested in the changing frequency of climate words.
- Bentley, R. Alexander, Philip Garnett, Michael J. O’Brien, and William A. Brock. 2012. “Word Diffusion and Climate Science.” PloS One 7 (11): e47966.



Boom and Bust of Science Dissemination...

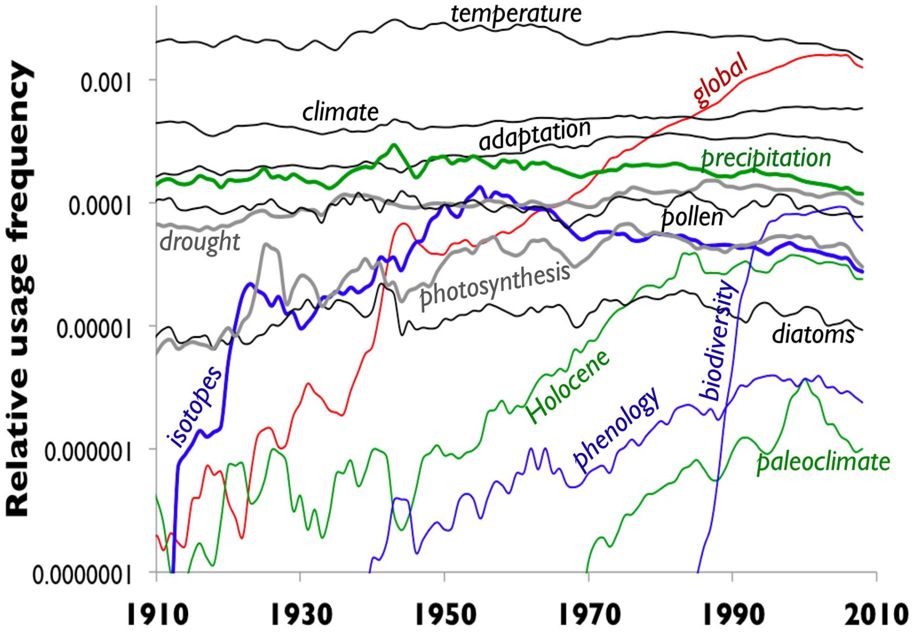
We were interested:

- In the cycles of interest in science.
- The role science dissemination has in ‘public interest’.
- Is it reflected in word frequencies?
- Is there a social aspect to the change in use of some words?

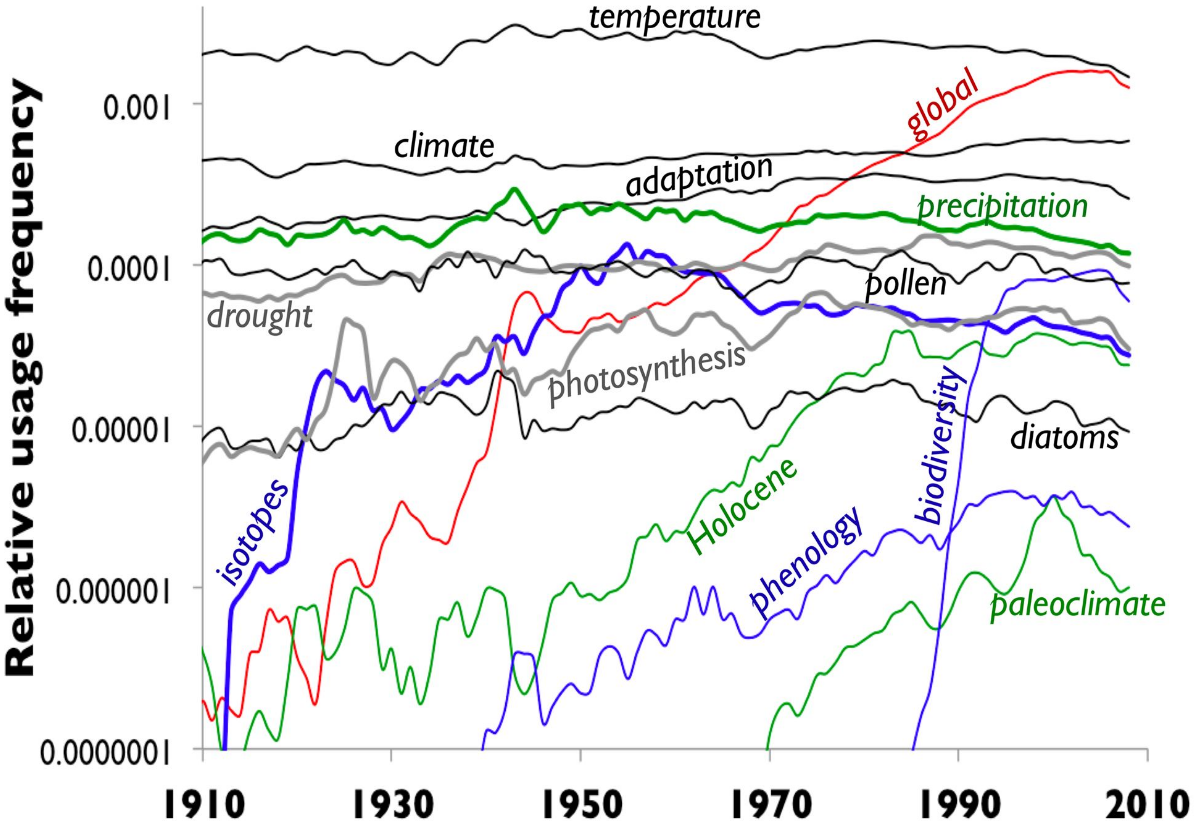


(modded) Bass Diffusion Model

- Models adoption of ‘products’ in a population…
- Probability of a word being used is proportional to the amount of times it is eventually used, and the rate of independant discovery. (so we are looking backwards).
- Predicts that things follow a s-shaped curve that in part depends on adoption by new



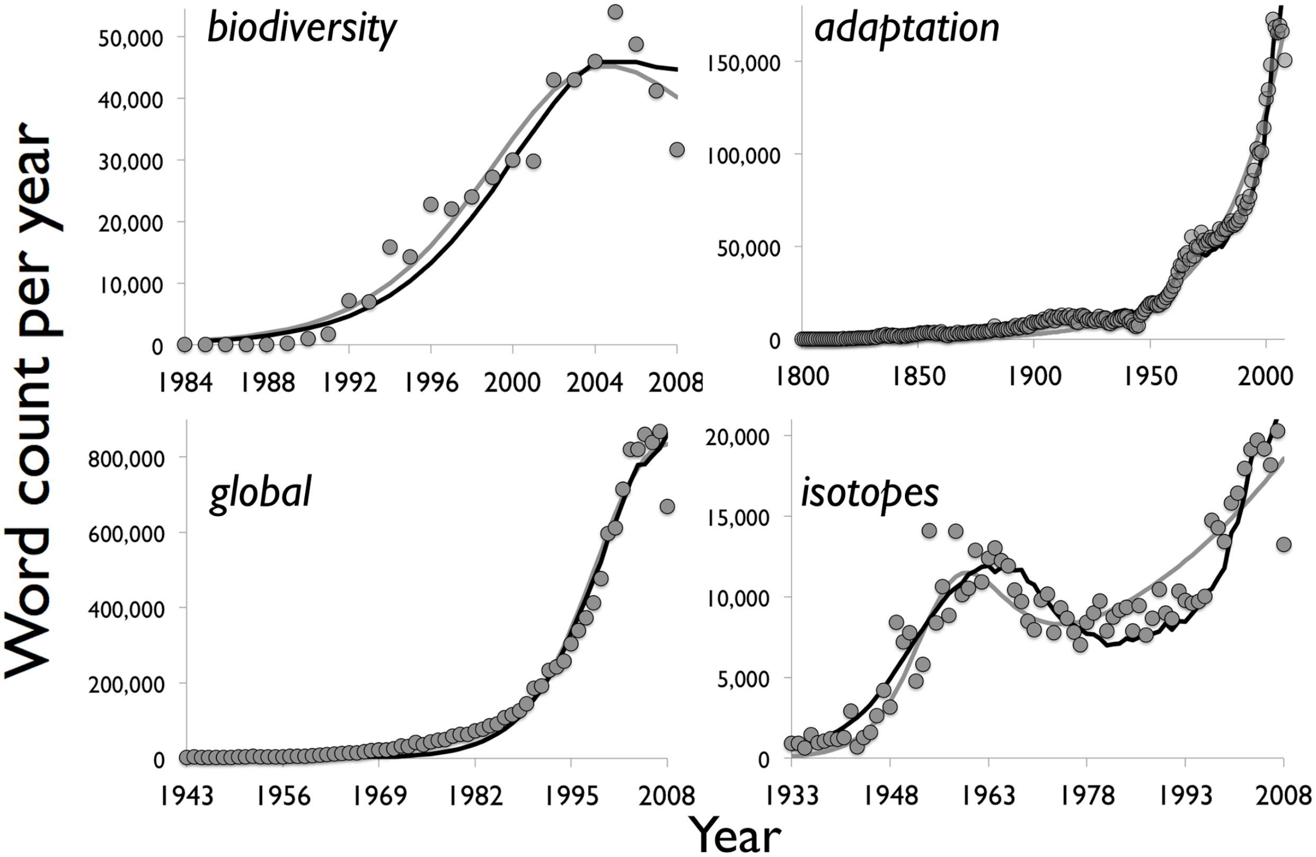
Results



Some follow modified bass model:
Biodiversity
Paleoclimate
Global
Holocene

Some don’t:
Temperature
Climate
Diatoms

Suggesting that the usage of some words is more influenced by social factors that others.



We then found that if we normalized to the frequency of *the* instead of the total number of ngrams per year we got better fits for the model to the data.

For the period *the* is the word with the highest frequency in the English language.

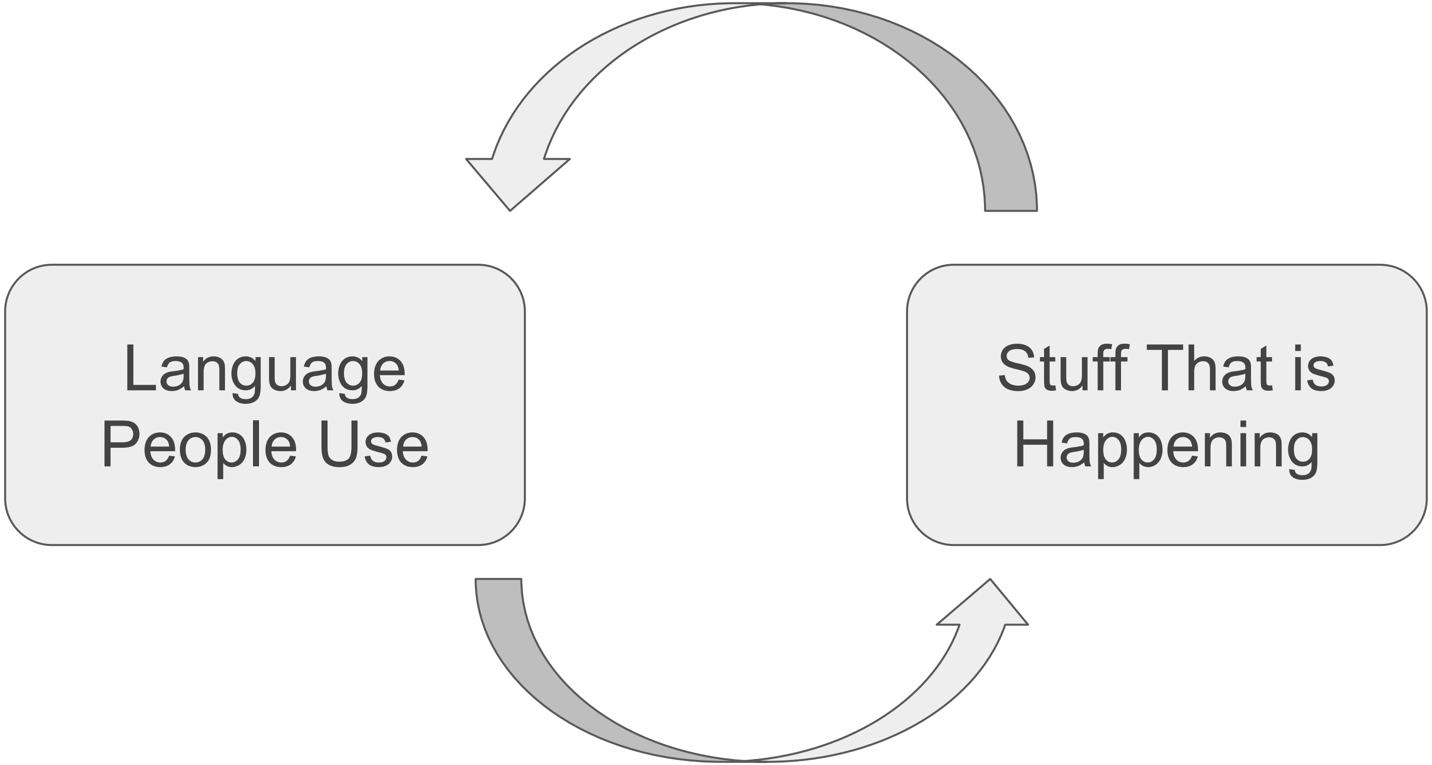
This is actually the case for most of the data.

Conclusion

Some of the words associated with climate change are changing in frequency due to some sort of social-learning diffusion process.

- So its not *selection* is *diffusion (drift)* - shift due to people copying other people.
- Or is it… so this is where it is useful to introduce the idea that when modelling/simulating something you first try the simplest model and then go from there.
- We are unable to reject the hypothesis that it is diffusion and not selection.

Emotion Words!



Language People Use

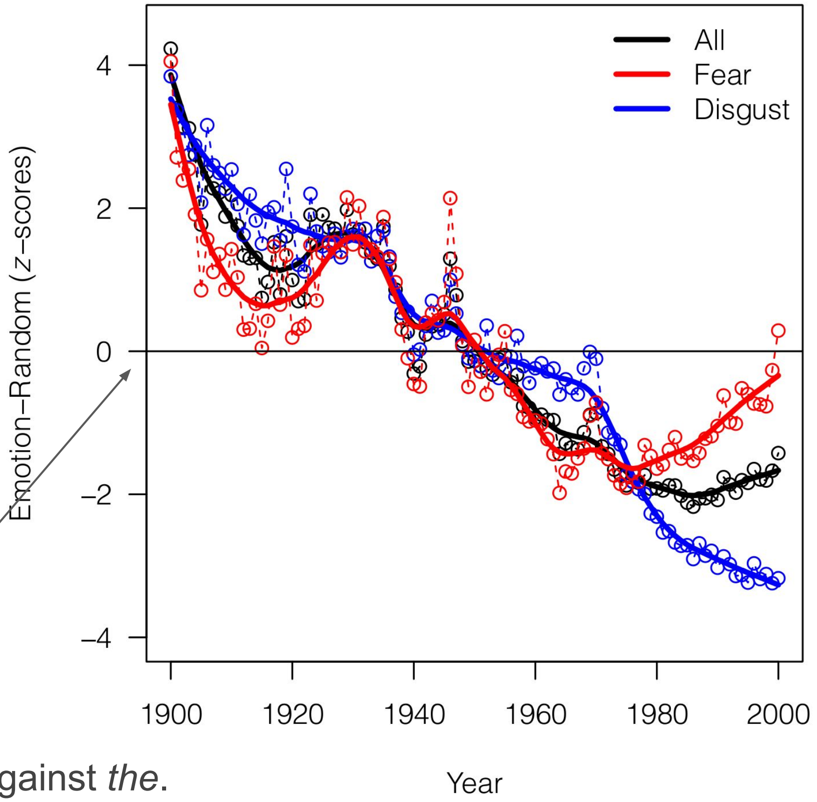
Stuff That is Happening

Feedback between language and the *World*.

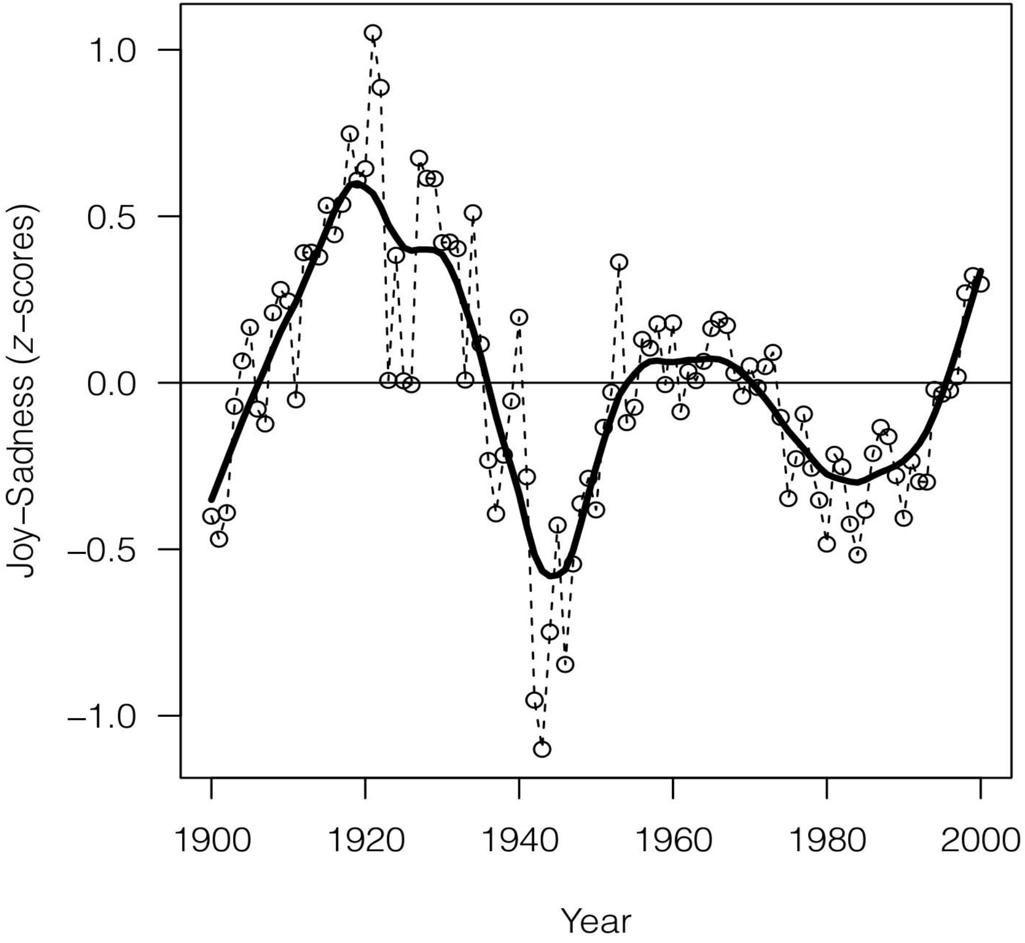
Emotion Words

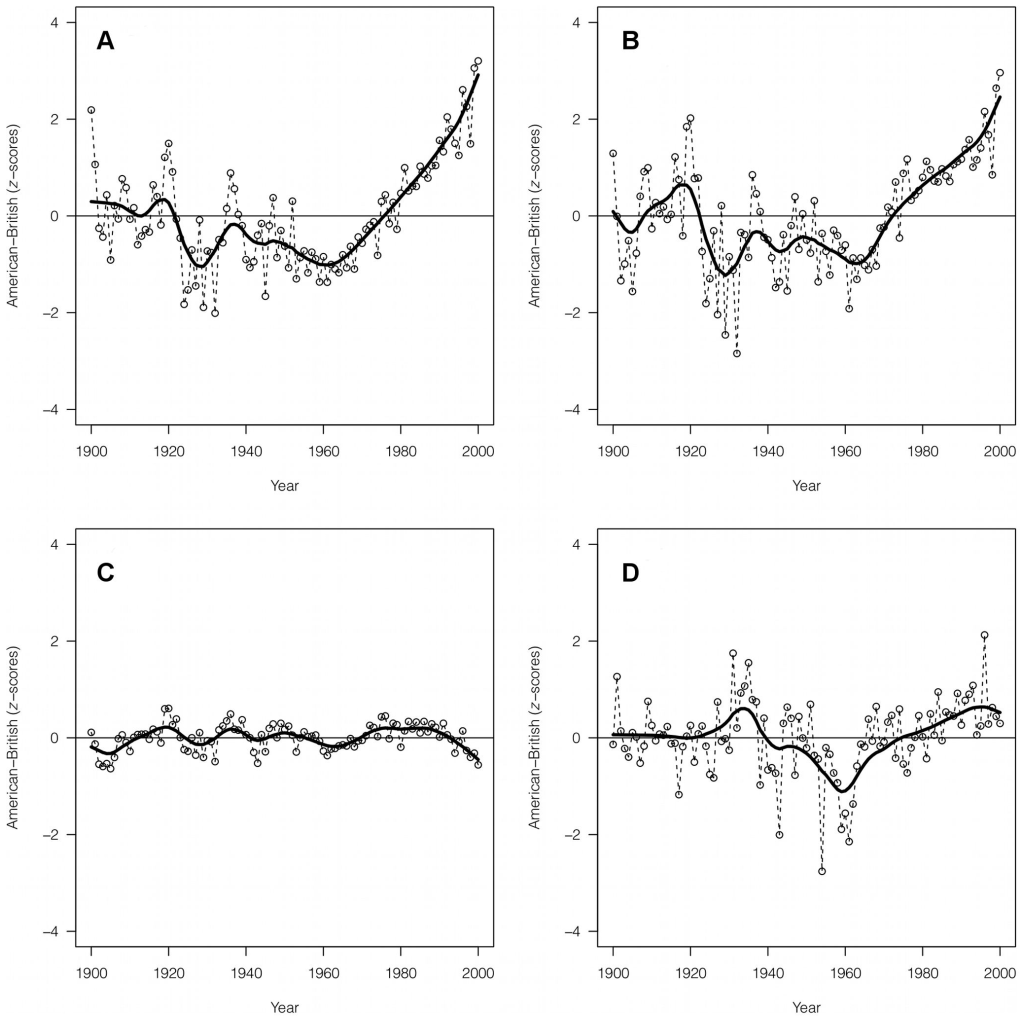
Does what is happening manifest in word frequencies?

- One would assume so…
- Can we see it?
- Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. “The Expression of Emotions in 20th Century Books.” PloS One 8 (3): e59030.



These are residuals (z-score) against *the*. Emotions words going down… except fear in the 2000s.


Joy–Sadness (z–scores)
1.0
0.5
0.0
–0.5
–1.0
1900
1920
1940
1960
1980
2000
Year



A - Emotion terms - American books have increased their mood terms.

B - Content free terms - America is more ‘content-free’? Stylistic change between US and British books.

C - Random (control).

D - 100 largest urban agglomerations in the world. Use of the names of the 100 most populated cities (control).

Conclusions

You can see evidence of a connection between language and *stuff happening*.

The exact relationship is perhaps not clear (mechanism etc).

Some things to note about this data...

- What is an emotion word? **(we tend to use other researcher’s data)**
- We have to assume that the division between British and American books is robust enough for the analysis (we don’t know exactly what is in each data set).

Other Studies

Work on depression that is interesting. There is a indication that the language used by people suffering depression shows distinct word frequency differences when compared with non-sufferers.

- What is that feedback system like??
- Al-Mosaiwi, Mohammed, and Tom Johnstone. 2018. “In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation.” Clinical Psychological Science 6 (4): 529–42.

We also looked at changes in individualism vs collectivism words in 20th century Russian… showed some interesting things but wasn’t as good as we hoped.

- Skrebyte, Agne, Philip Garnett, and Jeremy R. Kendal. 2016. “Temporal Relationships Between Individualism–Collectivism and the Economy in Soviet Russia: A Word Frequency Analysis Using the Google Ngram Corpus.” Journal of Cross-Cultural Psychology 47 (9): 1217–35.

The Neutral Model

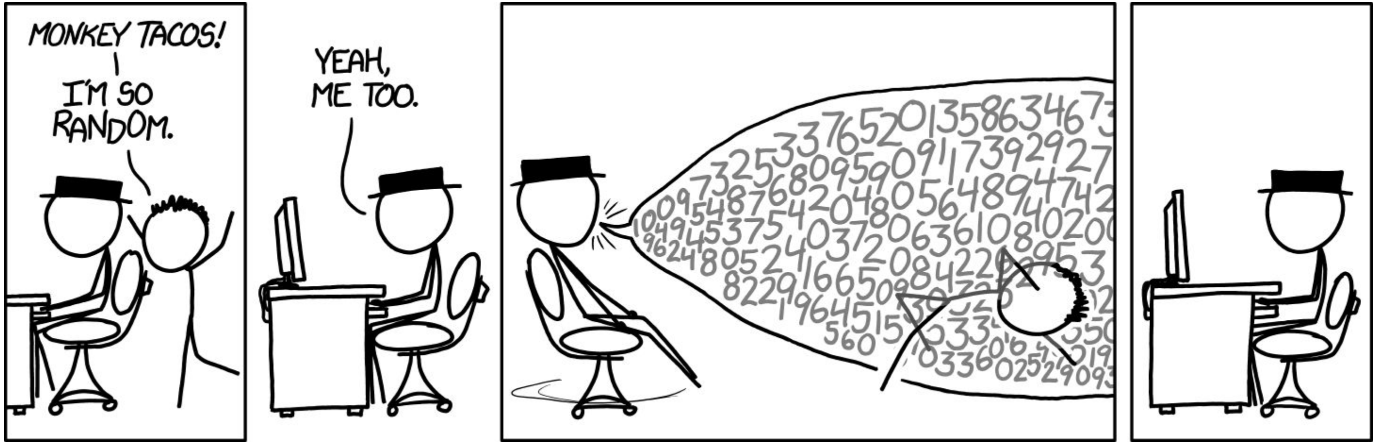
Neutral Model!

Neutral model (annoys some people):

- Assumes that things change via a neutral copying process.
- Thus your probability of doing something is proportional to the probability that you observe it and copy the behaviour, with an innovation rate (prob that you just invent something new).
- Therefore the *fitness* (its neutral!) of the choice doesn’t really matter (source of annoyance).
- Similar to genetic drift (oh dear).
- Seems to fit a lot of human behaviour (also source of annoyance).

Neutral Model and Words

- Word frequencies changes fit the neutral model - various forms of it too. Modified Bass earlier and also Neutral model just described.
- The words you use are more about whether you observe and copy and not so much about the word itself?
- Seems that we aren’t able to reject the neutral model in this case.



Neutral Model and Words

So similar problem to genetics… where does the selection come in and how can you tell the difference between word frequencies changing by selection and words changing by neutral processes?

- Probs need more data? Maybe not as genetics has tons of data and they still struggle.
- What about forgetting words?

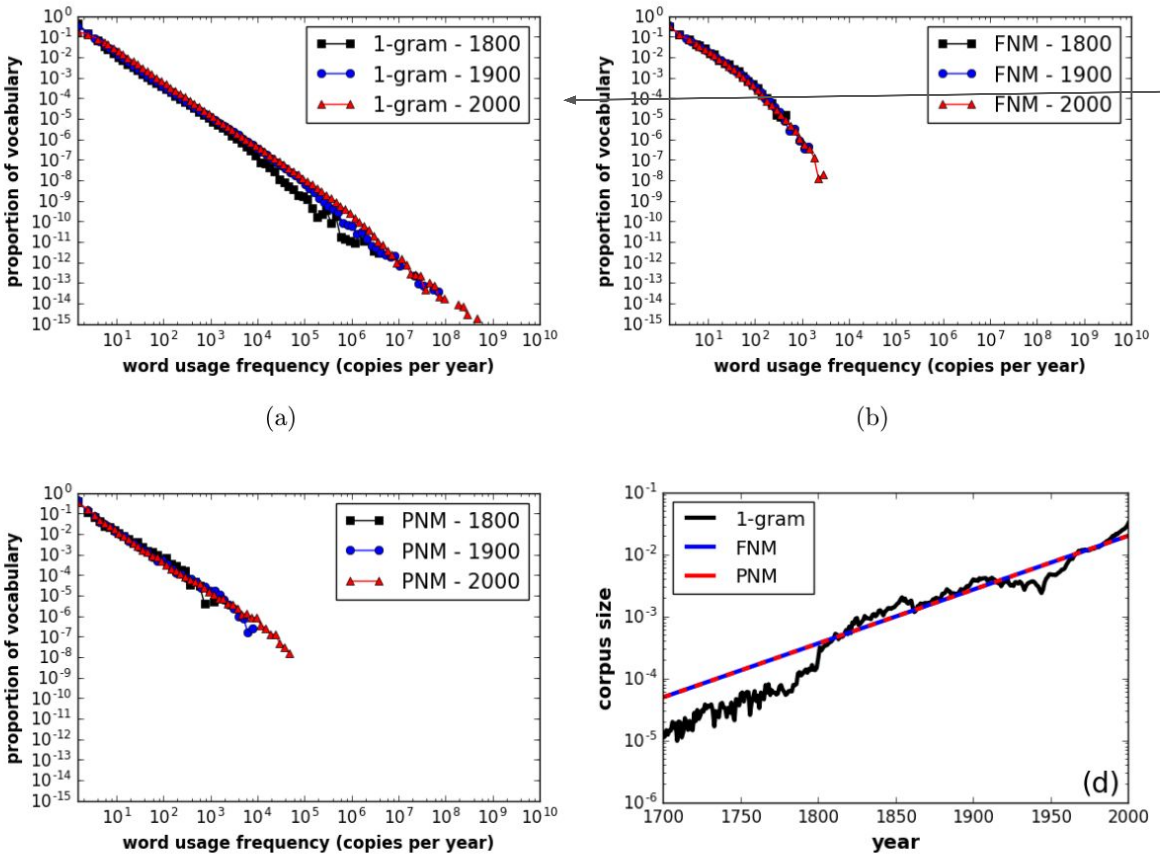
Drifting Forever?

- There seems to be a difference when it comes to forgetting words.
    - Or rather words declining in frequencies and drifting out of use.
    - Turnover!
- Perhaps it happens quicker than it should?
- That weird because it would imply that somehow we *know* when a word is declining in frequency.
- Then select to not use it?

So We Wrote a Paper on Turnover

...that no one reads.

- Interested in the ‘waxing and waning’ of the frequencies of words.
    - Including the turnover of words - relative rank in frequency.
    - E.g. position of a word in a topY by frequency.
- Two observations:
    - Zipf law holds for the ngram data set - the frequency of a word in inversely proportional to its frequency rank.
    - Heaps law holds, vocabulary size scales sub-linearly with total number of words.



(a)

(b)

(c)

(d)

Zipf’s Law, the frequency of a word in inversely proportional to its frequency rank

2 Neutral Models - FNM

FNM - full sampling neutral model

- “...would simply assume that authors choose to write words by copying those published in the past and occasionally inventing or introducing new words.”
- FNM reproduces Zipf’s law, and dynamic turnover is present.
- FNM predicts that a slowing down in the topY turnover as ‘new’ words find it more difficult to increase in in frequency by diffusion to get into topY (stochastic death).

FNM

Significantly FNM also, predicts that vocabulary will scale linearly with with the probability of the invention of words and the total number of words.

Does not match the data or Heaps law.

- Heaps law, vocabulary size scales sub-linearly with total number of words.

PNM

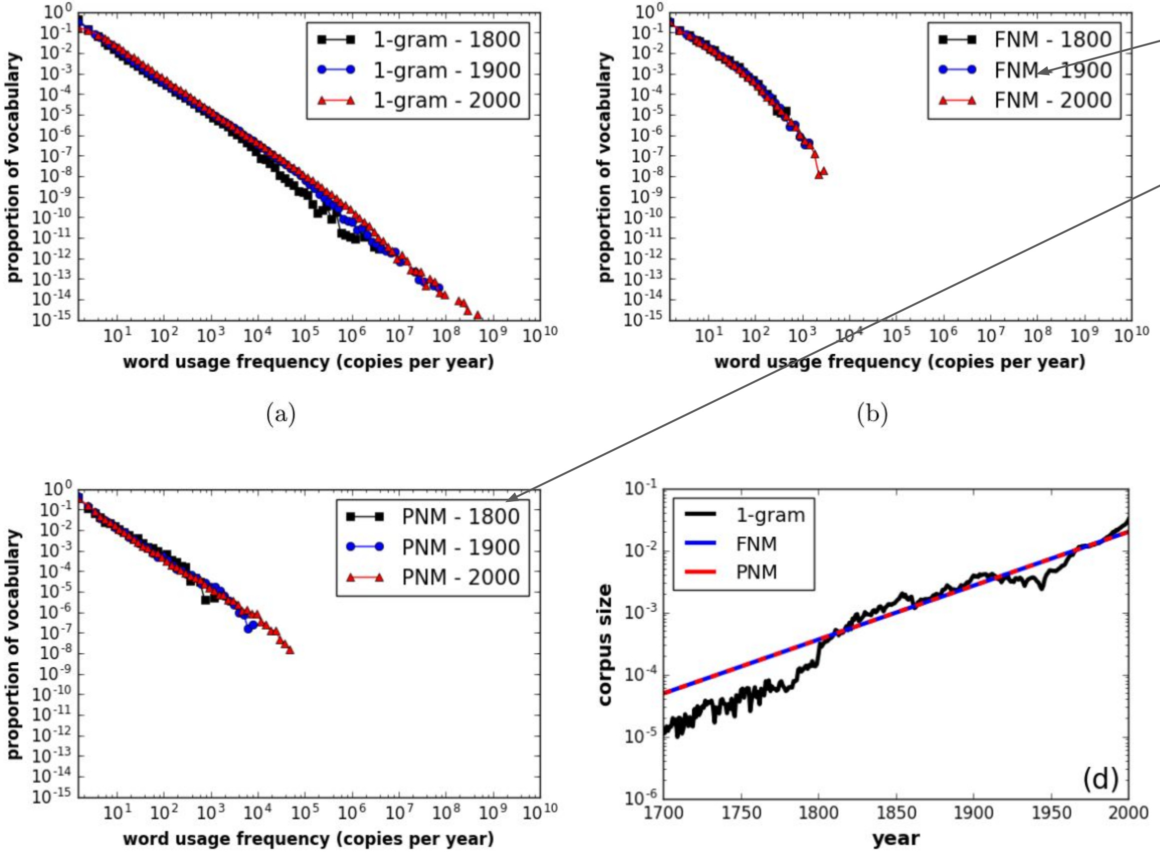
Ok, so hints in the data that the sub-linear scaling might be a recent invention.

So if we assume that and then assume that perhaps as the volume of available books increased that authors could only copy from a partial sample - an evolving subset or ‘canon’.

We get the partial-sampling Neutral model.

- *“There exists a an evolving small-subset of the world's books on which all writers are educated.”*

D. Ruck et al.



(a)

(b)

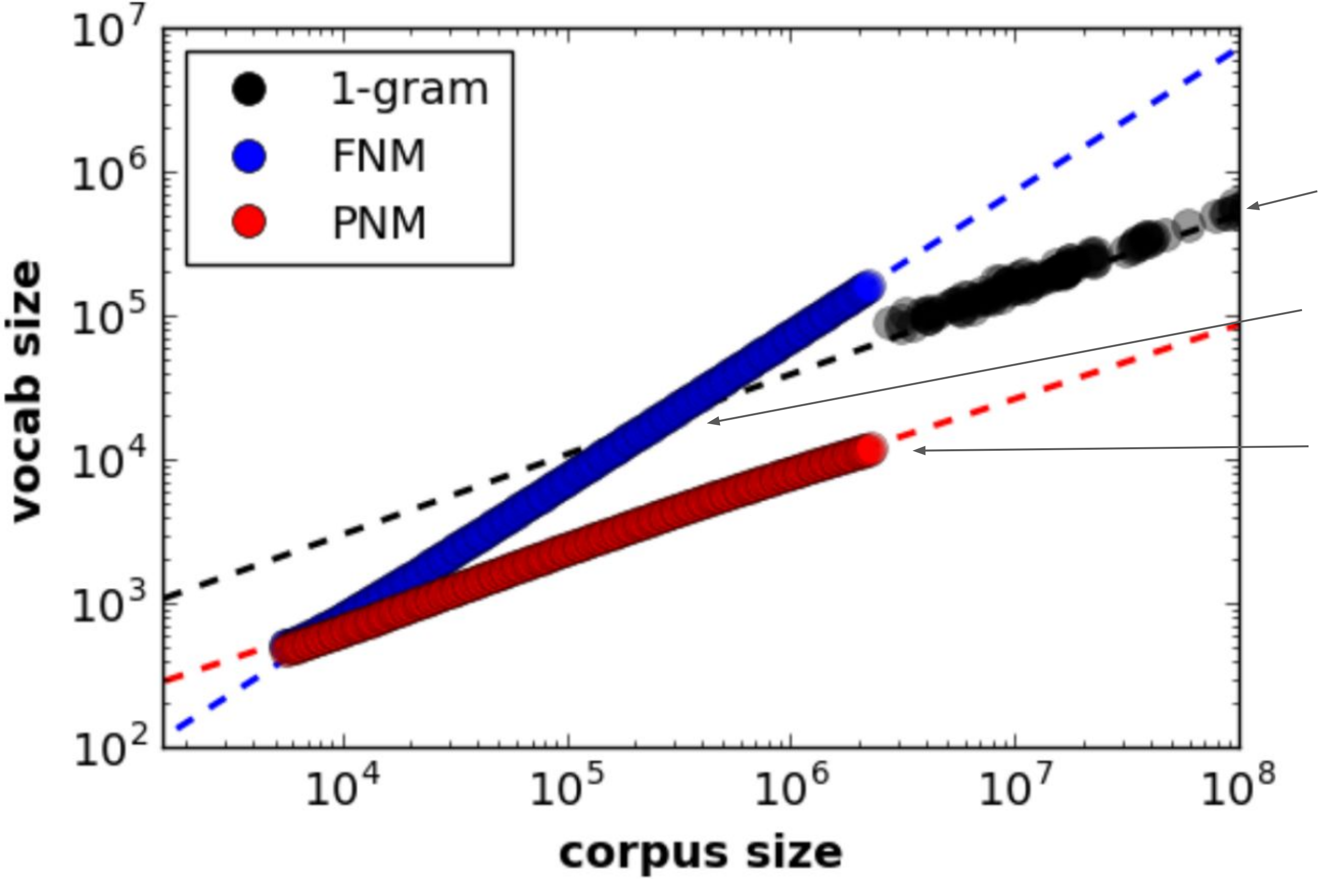
(c)

(d)

Both models fit Zipf's model.

PNM does it better for more orders or magnitude.

Computational limit stops us before we get to the same orders of magnitude as the data.



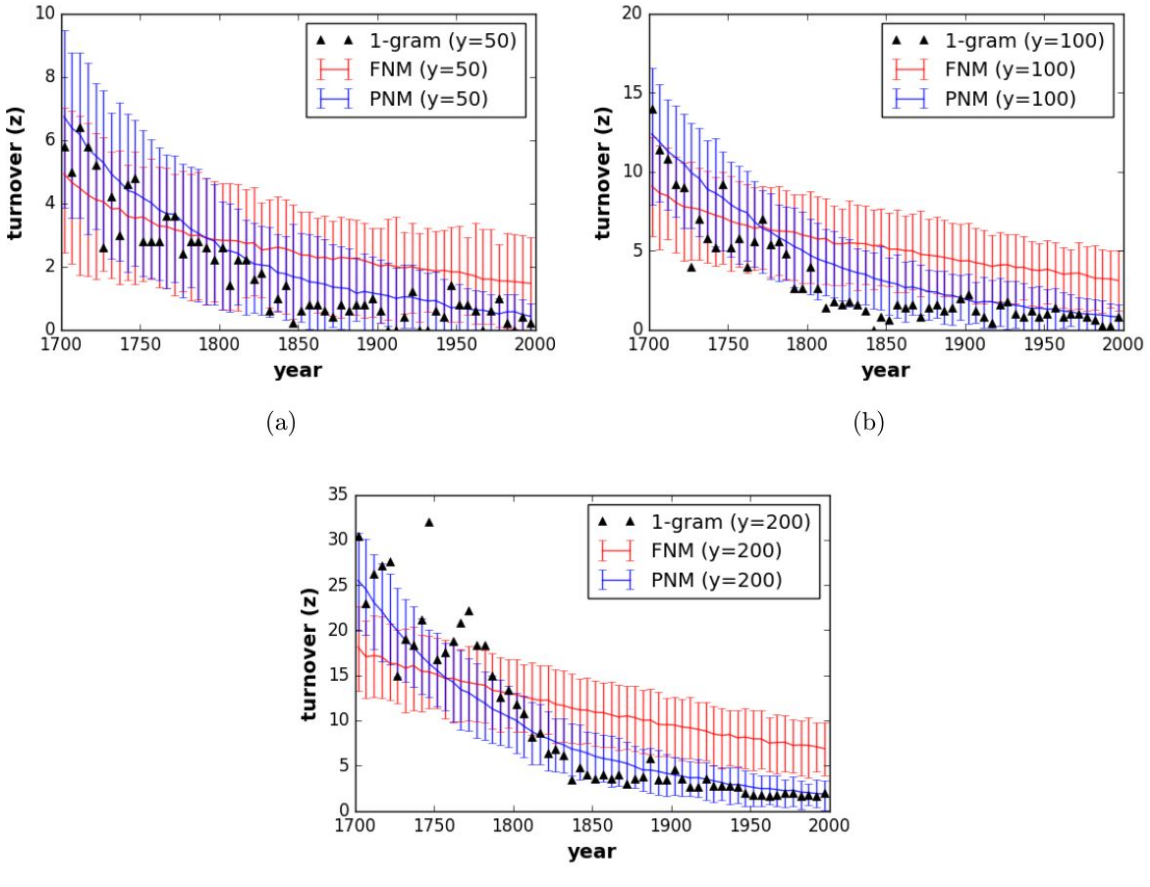
Heaps law:

Data scales sub-linearly (coefficient of less that 1).

FNM scales linearly with corpus size.

PNM scales sub-linearly with corpus size, same coefficient as data.

D. Ruck et al.



(a)

(b)

(c)

Turnover of topY

FNM never does very well turnover (esp if you are going to get Zipf right as well).

PNM does fairly well in the top 50 and 100 words, but starts to break down with the top 200 suggesting that the model might not be exactly correct.

So we can reject the FNM model, and perhaps partially reject the PNM model but we are not as clear on that one...

Conclusions

There is some sort of relationship between what words we use and *what* is happening. This is some sort of feedback system…

Neutral model(s) in various forms do model some aspects of how word usage evolves, providing a basis for the testing of hypotheses.

The PNM model suggests a plausible model for why words drop out of use more quickly than the FNM would predict (they are lost from the ‘canon’).