# Using big data to quantify the evolution of written corpora at the micro and macro scale

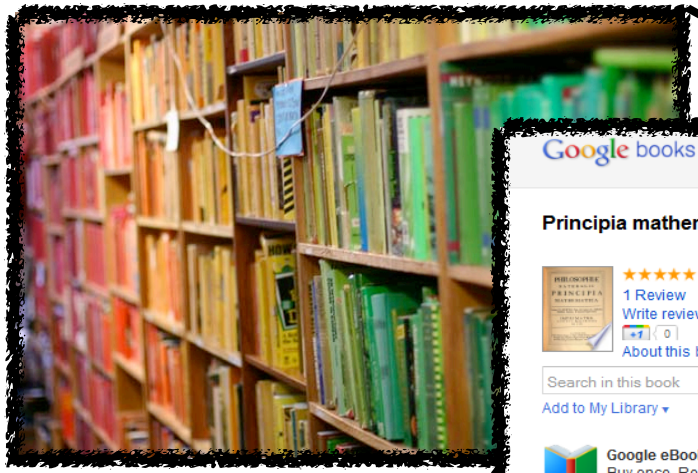
Alexander M. Petersen

IMT Lucca Institute for Advanced Studies, Lucca 55100, Italy

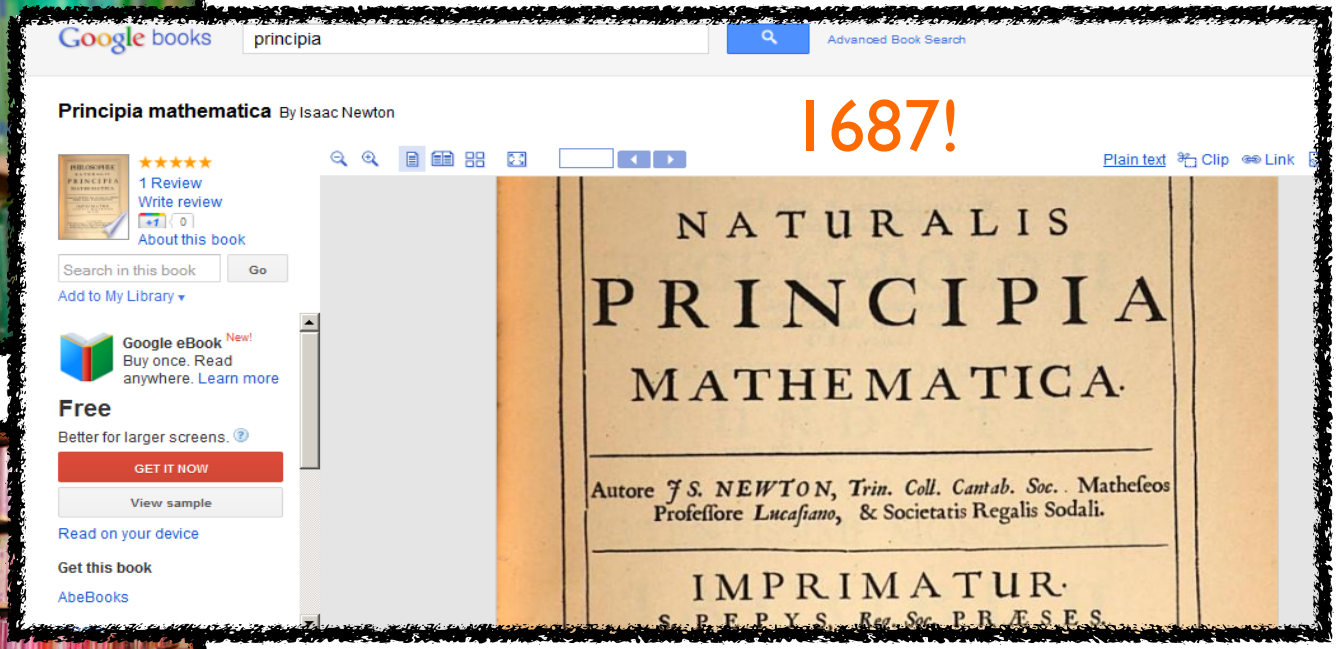
# *Outline*

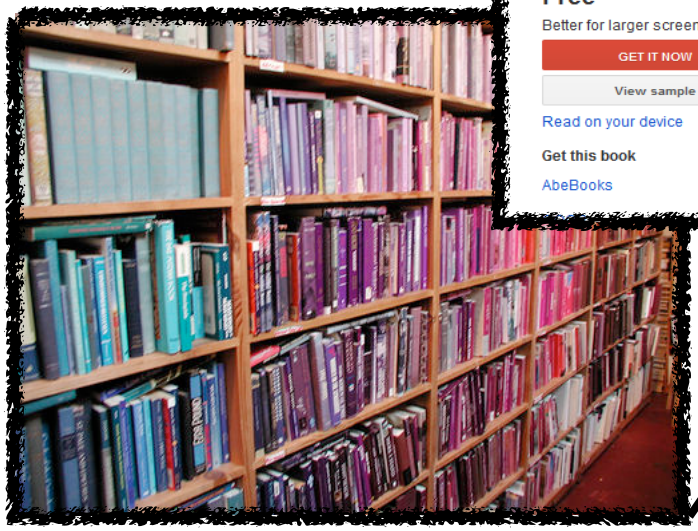
- Digital Humanities made possible by crowd-sourced "big data" spanning multiple levels of order (time, geography, ...)
  - Books: n-grams ~ word space
  - Journal articles: technical terms ~ idea space (MeSH)
- Quantitative analysis of historical trends
  - Competition (e.g. for limited attention)
  - Geospatial variation and the role of socio-political shocks
  - Growth trends in the use of individual (new) words
  - Zipf's law and Heaps' allometric scaling over time

# *Historical crowd-sourced data*



*Google Inc.* digital books repository



Corpus of 5,195,769 digitized books from 1520-present, containing ~4% of all books ever published

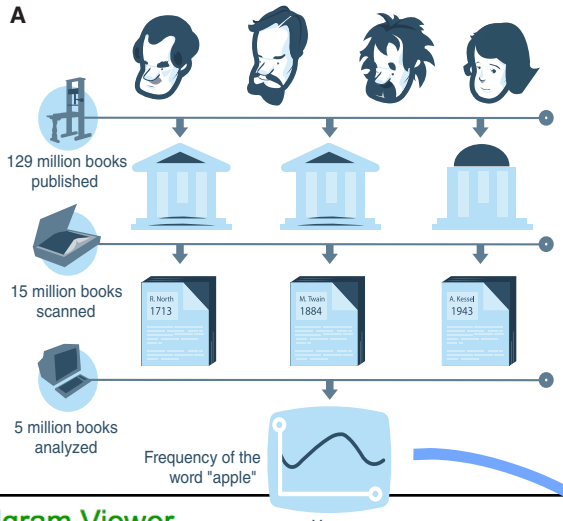
**Quantitative Analysis of Culture Using Millions of Digitized Books**

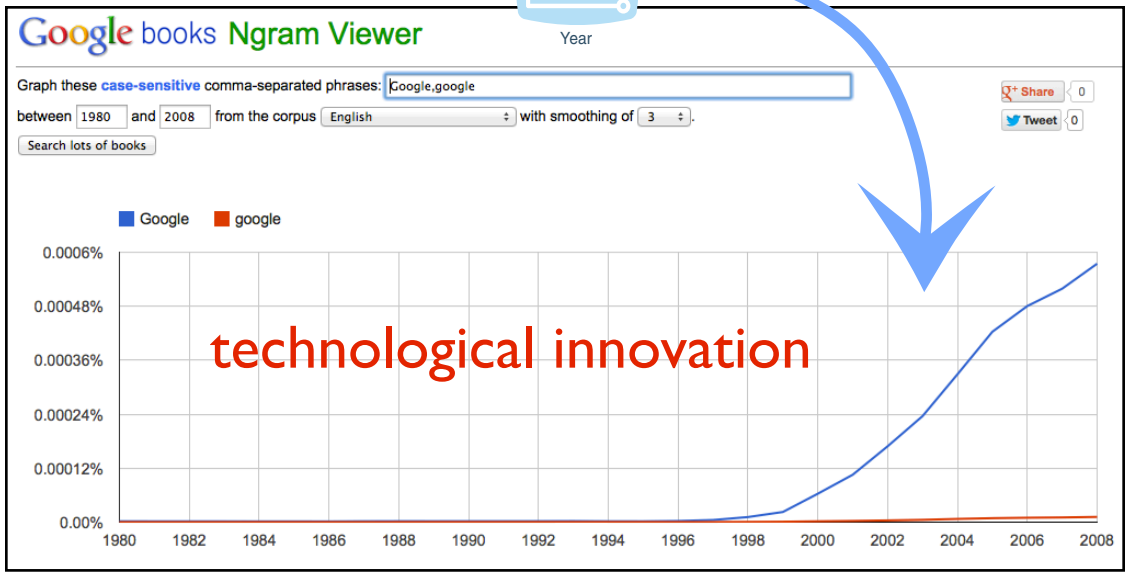
14 JANUARY 2011 VOL 331 **SCIENCE**

# Time series constructed from word counts in books: aggregated at multiple levels

Michel, J.-B. et al. Quantitative analysis of culture using millions of digitized books. Science (2011).

*Google Inc.* digital books repository

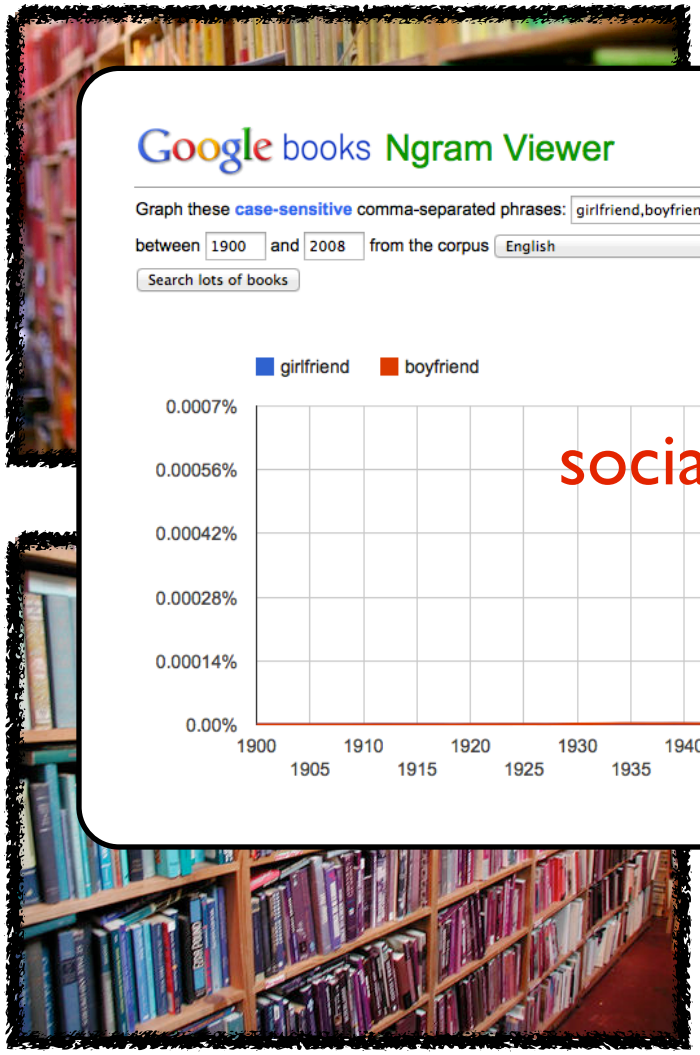


# Time series constructed from word counts in books: aggregated at multiple levels

Michel, J.-B. et al. Quantitative analysis of culture using millions of digitized books. Science (2011).

*Google Inc.* digital books repository



Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: girlfriend,boyfriend

between 1900 and 2008 from the corpus English with smoothing of 3.

Search lots of books

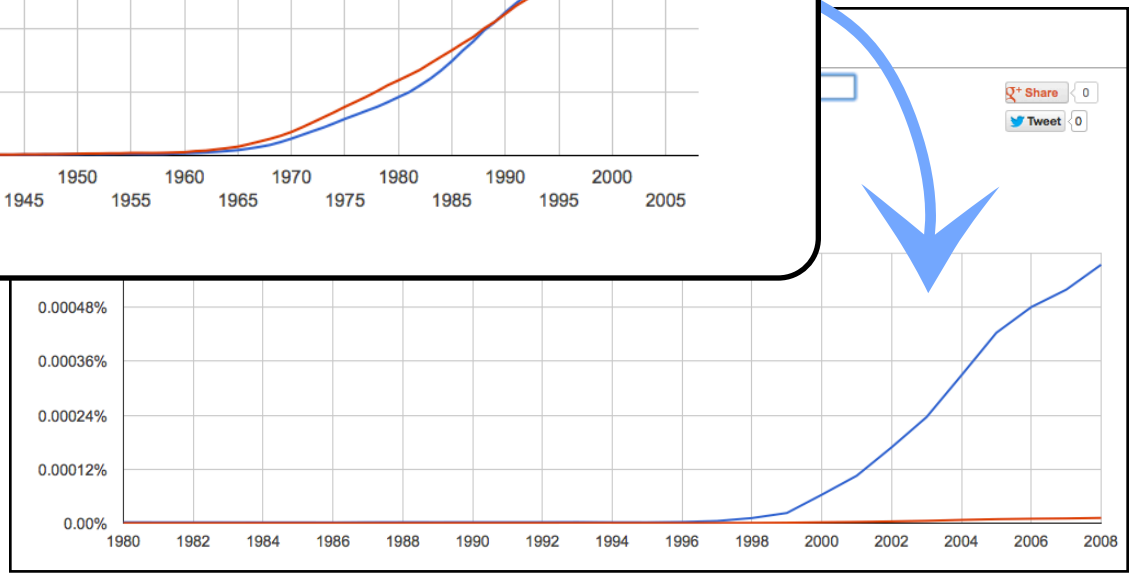
Share 0 Tweet 0

girlfriend boyfriend

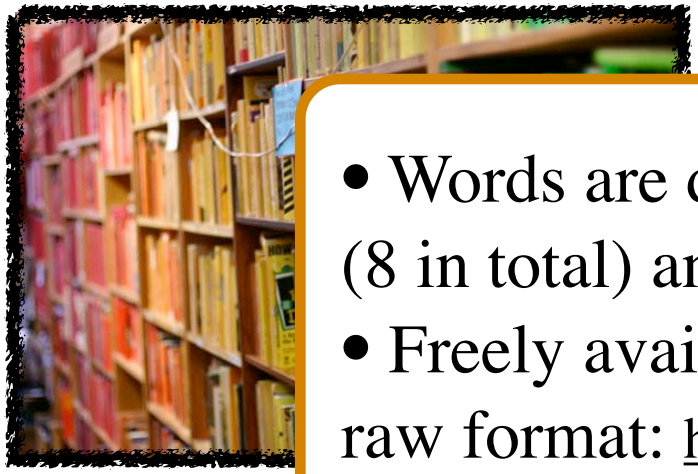
0.0007% 0.00056% 0.00042% 0.00028% 0.00014% 0.00%

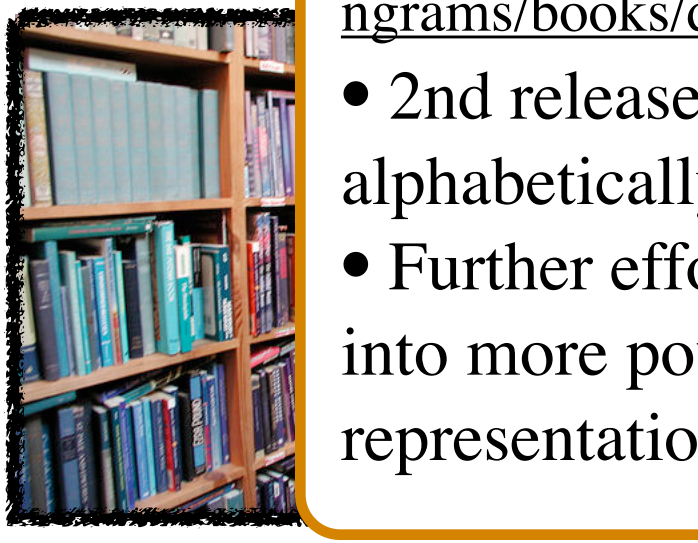
1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005

social innovation



# Time series constructed from word counts in books

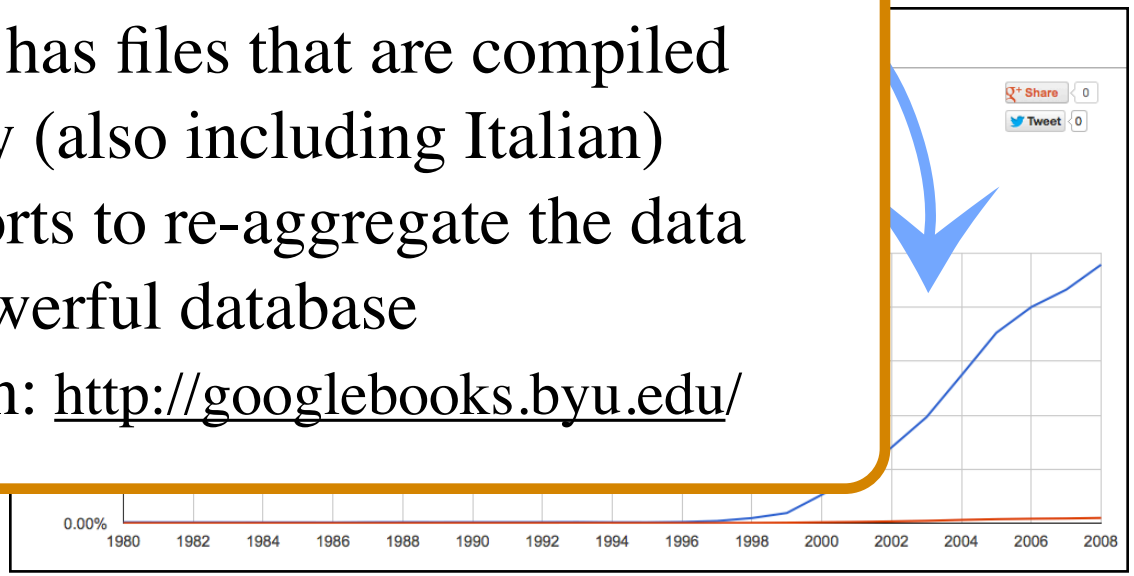




Michel, J.-B. et al. Quantitative analysis of culture using millions of digitized books. Science (2011).

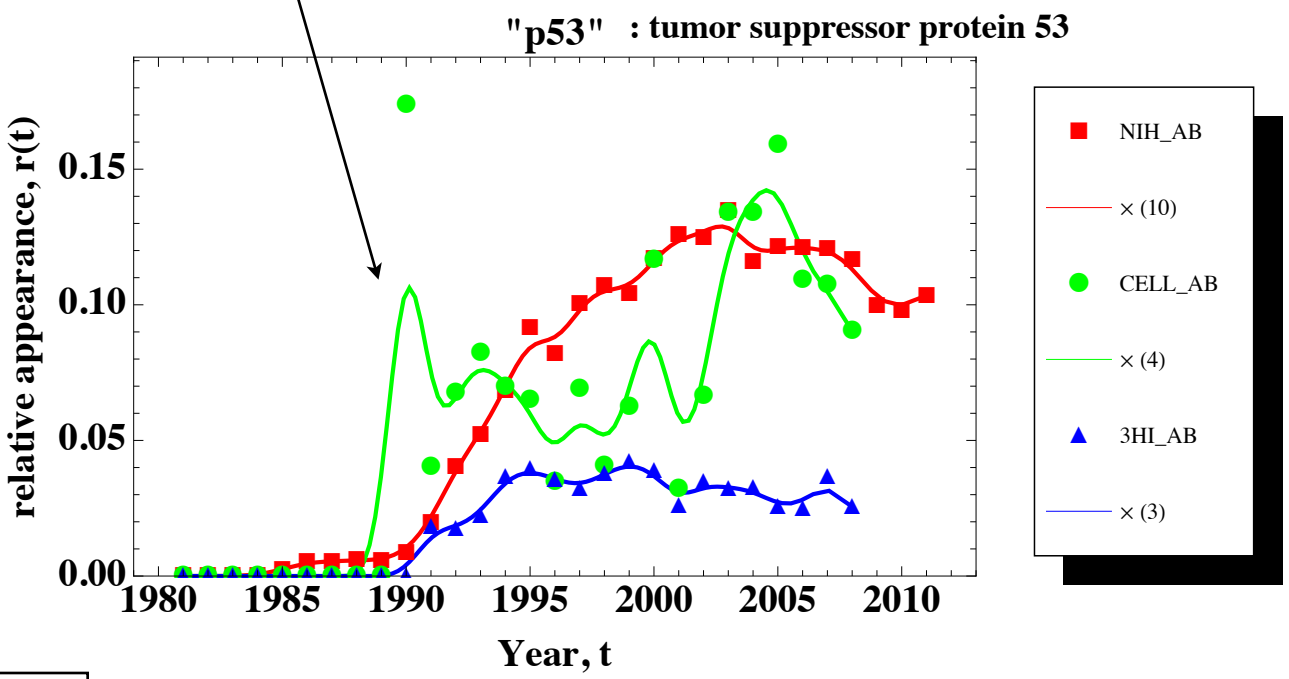
*Google Inc.* digital books repository

- Words are disaggregated across language (8 in total) and by word/page/book count
- Freely available & easy to download in raw format: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
- 2nd release has files that are compiled alphabetically (also including Italian)
- Further efforts to re-aggregate the data into more powerful database representation: http://googlebooks.byu.edu/



# Time series constructed from word counts in journal abstracts and titles

Baker SJ, et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244, 217–21 (1989)




NIH National Institutes of Health
Turning Discovery Into Health

nature

Cell

Science

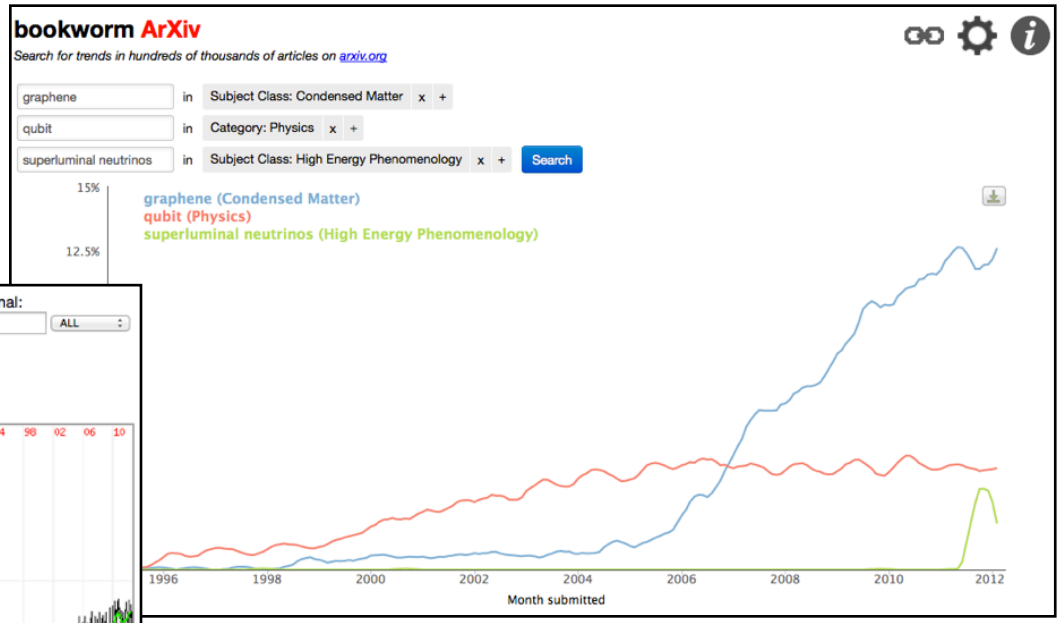
AAAS

The NEW ENGLAND JOURNAL of MEDICINE

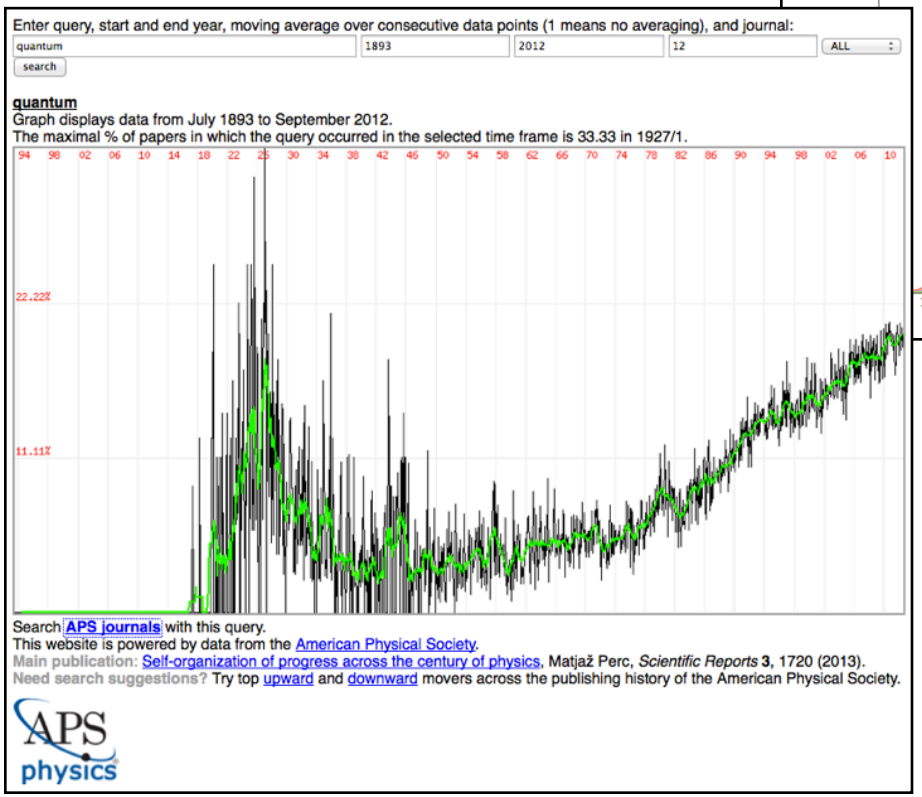
PNAS
Proceedings of the National Academy of Sciences of the United States of America

# Time series constructed from word counts in ArXiv articles and Physical Review articles

http://arxiv.culturomics.org/




http://www.matjazperc.com/aps/physics.html



# *Beyond text to context*

words: context free (e.g. cold = temperature, sentiment, sickness)
idea space: MeSH = controlled thesaurus of subject headings


**MeSH Terms**

- Alleles
- Animals
- Chromosome Deletion*
- Chromosomes, Human, Pair 17*/ultrastructure
- Colorectal Neoplasms/genetics*
- Humans
- Mice
- Mice, Nude
- Mutation*
- Neoplasm Proteins/genetics*
- Nucleic Acid Hybridization
- Oncogenes
- Phosphoproteins/genetics*
- Suppression, Genetic
- Tumor Suppressor Protein p53

# What is meant by “mice”?

**2013 MeSH**

**MeSH Descriptor Data**

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

| | |
|---|---|
| **MeSH Heading** | Mice |
| **Tree Number** | B01.050.150.900.649.865.635.505.500 |
| **Annotation** | check tag: NIM no qualifiers for MICE, the genus MUS unspecified, or any MUS species |
| **Scope Note** | The common name for the genus Mus. |
| **Entry Term** | Mice, House |
| **Entry Term** | Mice, Laboratory |
| **Entry Term** | Mouse |
| **Entry Term** | Mouse, House |
| **Entry Term** | Mouse, Laboratory |
| **Entry Term** | Mouse, Swiss |
| **Entry Term** | Mus |
| **Entry Term** | Mus domesticus |
| **Entry Term** | Mus musculus |
| **Entry Term** | Mus musculus domesticus |
| **Entry Term** | Swiss Mice |
| **Allowable Qualifiers** | AB AH BL CF CL EM GD GE IM IN ME MI PH PS PX SU UR VI |
| **History Note** | 2006 |
| **Date of Entry** | 20050630 |
| **Unique ID** | D051379 |

# What is meant by “mice, nude”?

## National Library of Medicine - Medical Subject Headings

**2013 MeSH**

**MeSH Descriptor Data**

Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

| | |
|---|---|
| **MeSH Heading** | Mice, Nude |
| **Tree Number** | B01.050.150.900.649.865.635.505.500.550.500 |
| **Annotation** | NIM when exper animal: no qualif; when IM, qualif permitted; do not confuse with MICE, HAIRLESS see MICE, INBRED HRS; do not forget also to check tag MICE |
| **Scope Note** | Mutant mice homozygous for the recessive gene "nude" which fail to develop a thymus. They are useful in tumor studies and studies on immune responses. |
| **Entry Term** | Athymic Mice |
| **Entry Term** | Mice, Athymic |
| **Entry Term** | Mouse, Athymic |
| **Entry Term** | Mouse, Nude |
| **Entry Term** | Nude Mice |
| **Allowable Qualifiers** | AB AH BL CF CL EM GD GE IM IN ME MI PH PS PX SU UR VI |
| **Previous Indexing** | Mice (1966-1974) |
| **History Note** | 75 |
| **Date of Entry** | 19741111 |
| **Unique ID** | D008819 |

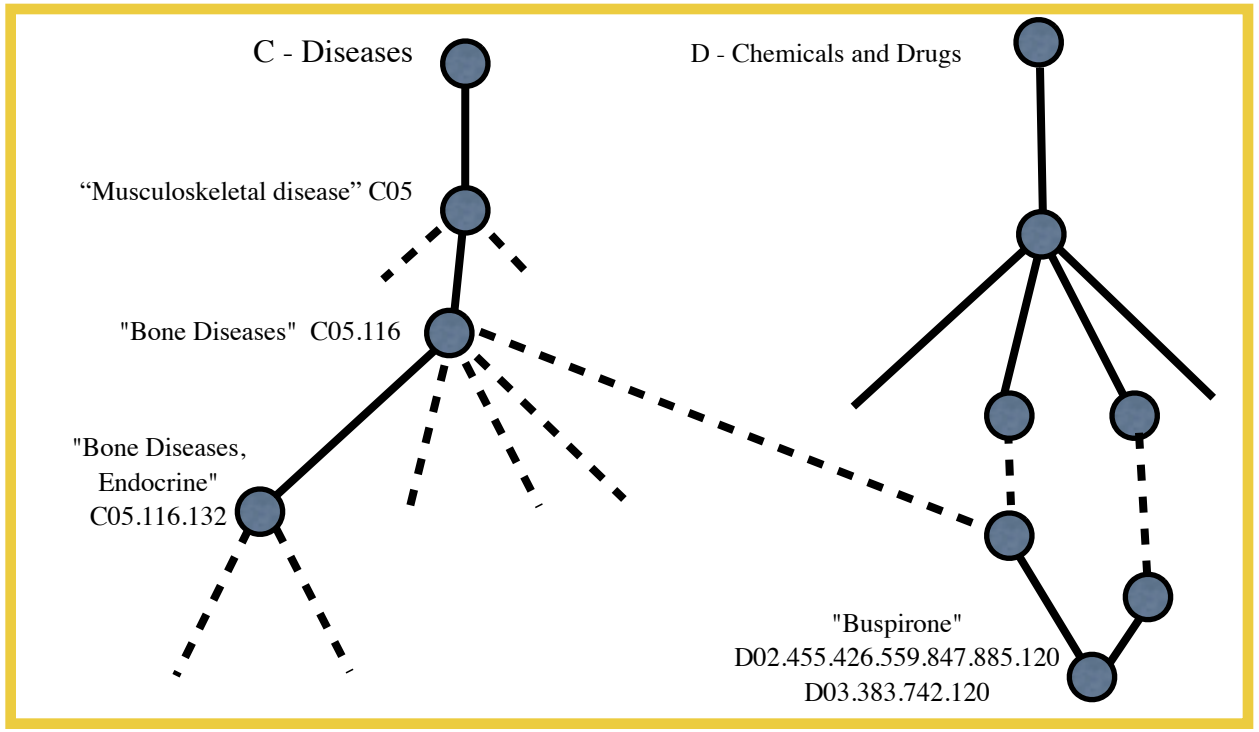
# *Knowledge Thesaurus*

**Medical Subject Headings (MeSH)** is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that facilitates searching. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings.

Mesh N-level tree Identifier: $(C\#\#)_1.(\#\#\#)_2...(\#\#\#)_N$

**Structure of MeSH**

The 2009 version of MeSH contains a total of 25,186 *subject headings*, also known as *descriptors*. Most of these are accompanied by a short description or definition, links to related descriptors, and a list of synonyms or very similar terms (known as *entry terms*). Because of these synonym lists, MeSH can also be viewed as a thesaurus.

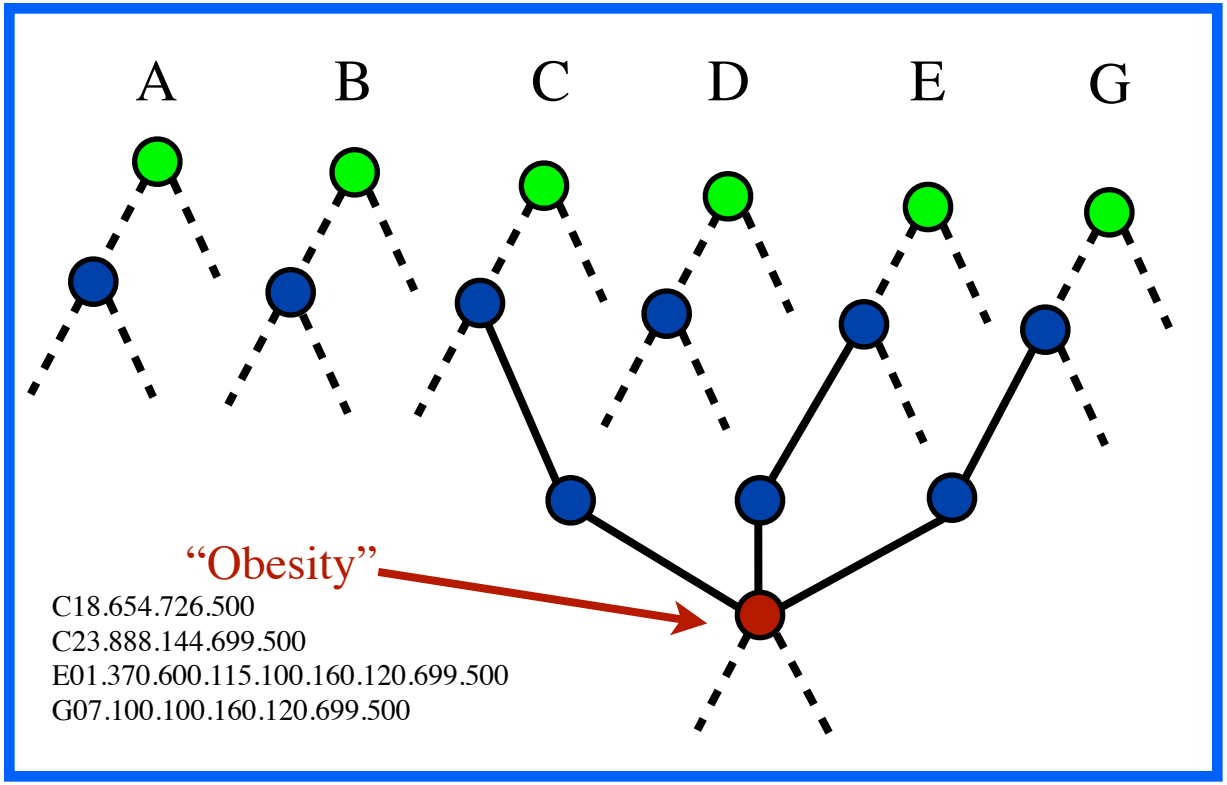


# *Spanning various thematic categories*

Each MeSH (Medical Subject Headings) term is a node in the MeSH tree with at least one tree identifier locating it at on branch $\alpha_i$ at level $N$:

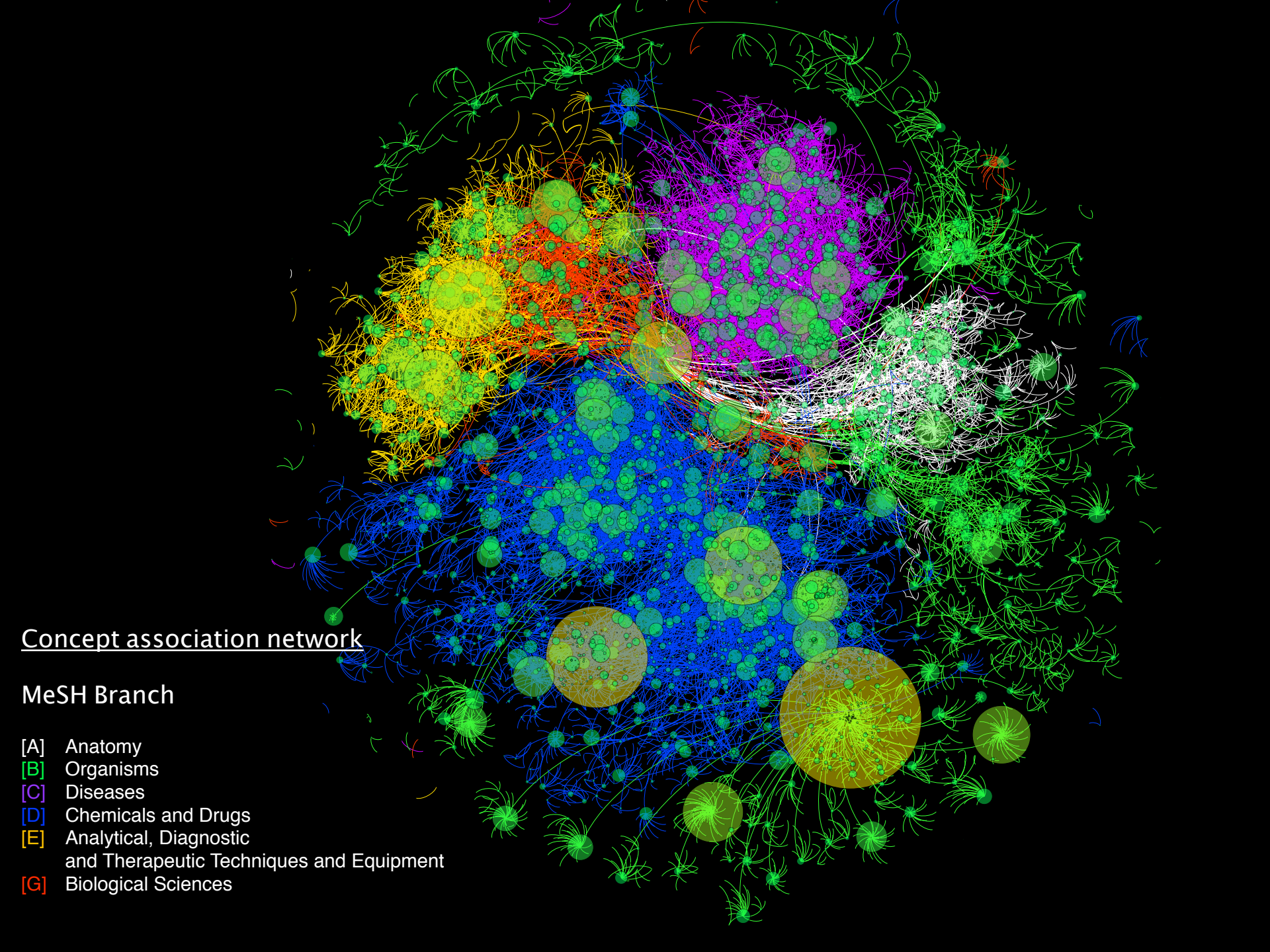
$$(\alpha_i\#\#)_1.(\#\#\#)_2...(\#\#\#)_N$$

Root Categories




A - Anatomy
B - Organisms
C - Diseases
D - Chemicals and Drugs
E - Analytical, Diagnostic and Therapeutic Techniques and Equipment
G - Biological Sciences

F - Psychiatry and Psychology
H - Physical Sciences
I - Anthropology, Education, Sociology and Social Phenomena
J - Technology and Food and Beverage
K - Humanities
L - Information Science
M - Persons
N - Health Care
V - Publication Characteristics
Z - Geographic Locations

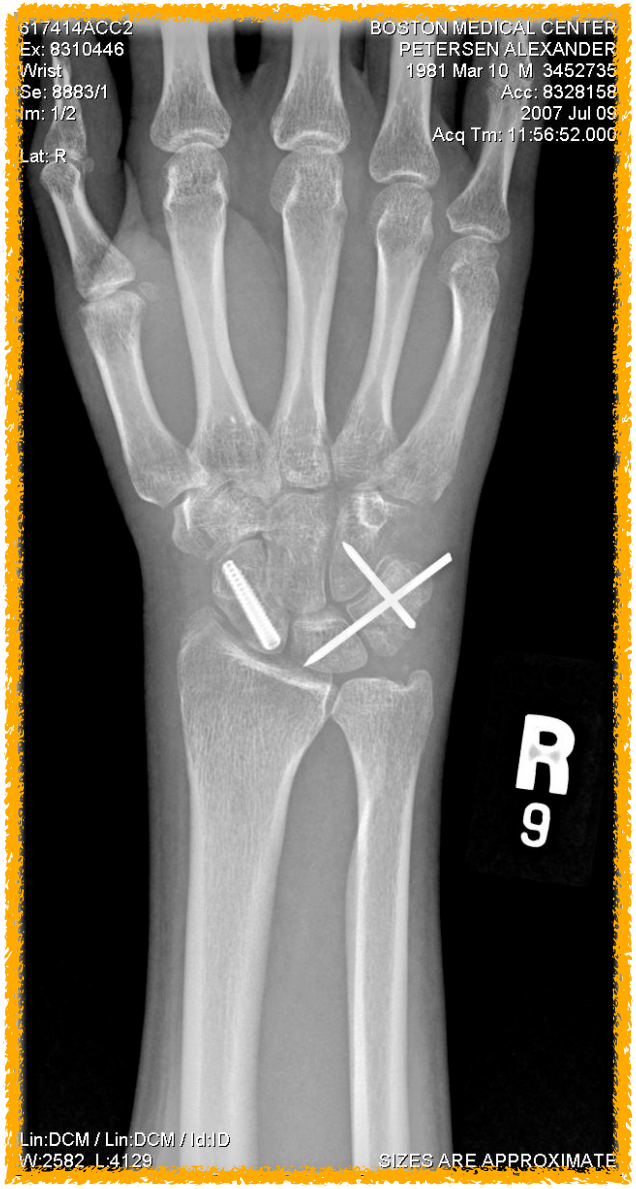

Concept association network
MeSH Branch
[A] Anatomy
[B] Organisms
[C] Diseases
[D] Chemicals and Drugs
[E] Analytical, Diagnostic
and Therapeutic Techniques and Equipment
[G] Biological Sciences

# *Language as a competitive system*



A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley.
**Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death**
Scientific Reports 2, 313 (2012).

# Do words compete in a linguistic "marketplace" for a finite market share?




Is this a:

a) Xray

b) Radiogram

c) Roentgenogram

??

# Competitive "marketplace"?

$u_i(t)$ : # of uses of word $i$ in year $t$

Total word usages in year $t$

$$N_u(t) \equiv \sum_{i=1}^{N_w(t)} u_i(t)$$

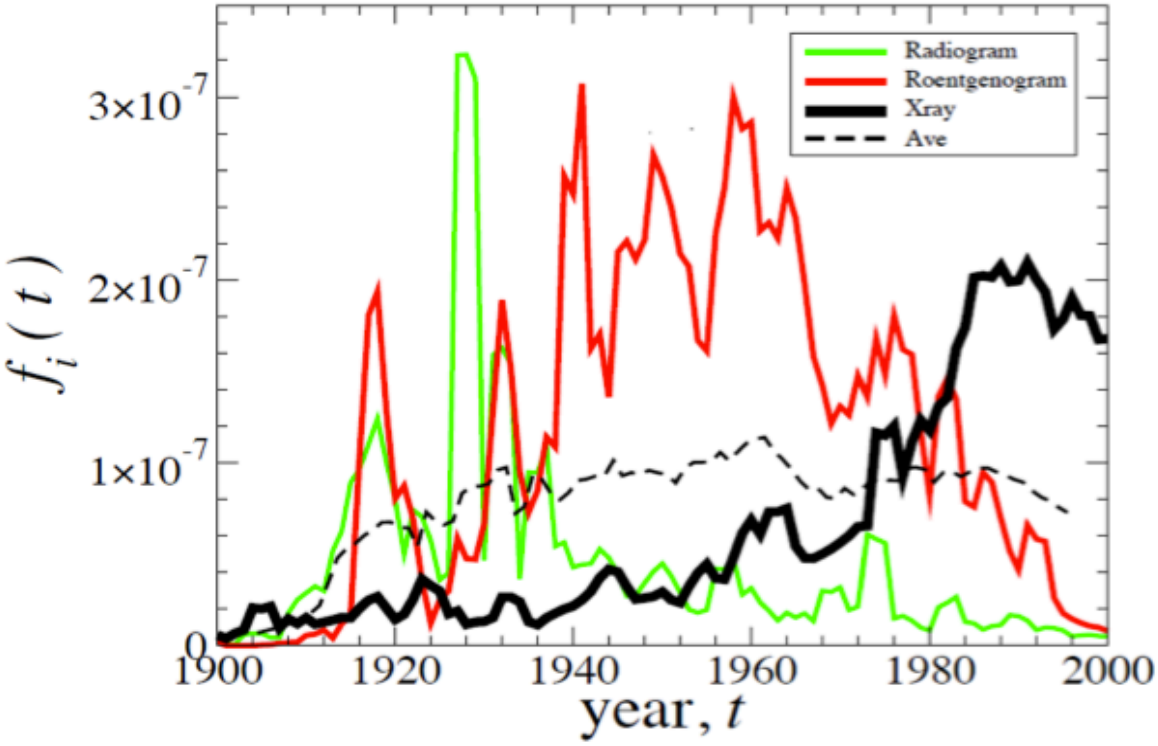
= corpus size

word frequency in year $t$

$$f_i(t) \equiv u_i(t)/N_u(t),$$

total number of distinct words in year $t$ :

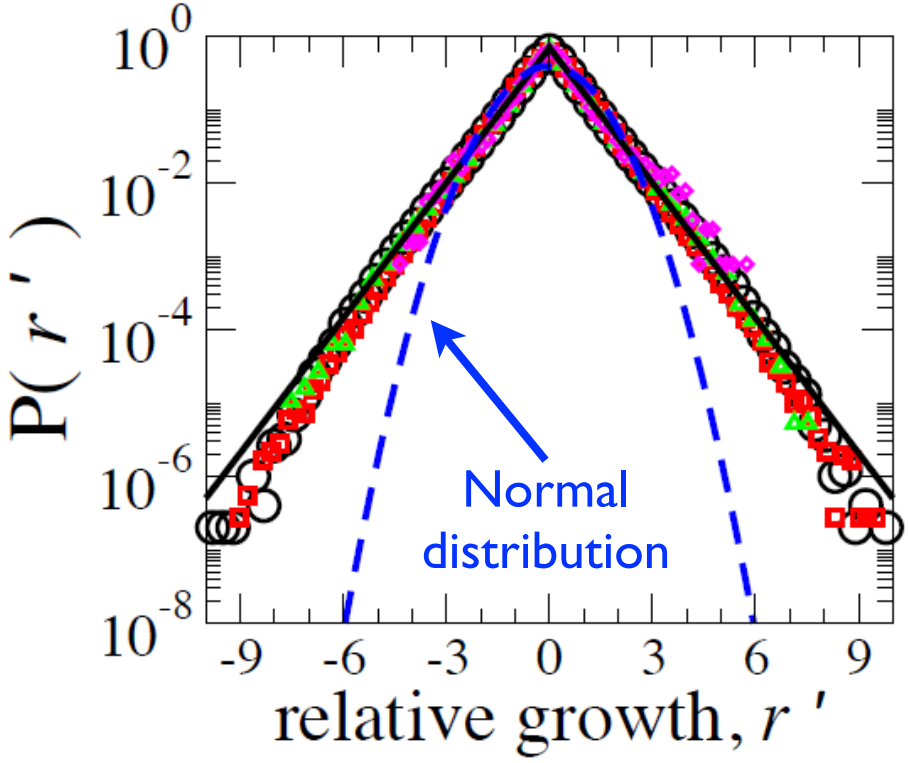
$$N_w(t)$$




Competition between:

- Synonyms
- Spellings (e.g. color vs. colour)
- other ideas in an abstract "idea space": (the Euphemism treadmill: **shell shock** (WWI) > **battle fatigue** (WWII) > **operational exhaustion** (Korean War) > **PTSD** (Vietnam War)

# Leptokurtic "tent-shaped" distribution of word usage growth rate




Common words
using $f_i \geq f_c$

| | | |
|---|---|---|
| ○ | English: | $f_c = 5 \times 10^{-8}$ |
| □ | Eng. (fict.): | $f_c = 10^{-7}$ |
| △ | Spanish: | $f_c = 10^{-6}$ |
| ◇ | Hebrew: | $f_c = 10^{-5}$ |

r = annual growth rates in the word usage frequency

$$r_i(t) \equiv \ln f_i(t+\Delta t) - \ln f_i(t) = \ln\left(\frac{f_i(t+\Delta t)}{f_i(t)}\right)$$

*P(r)* is centered around $r \cong 0$, a "zero sum" competitive system

# "tent-shaped" growth patterns are common in complex systems

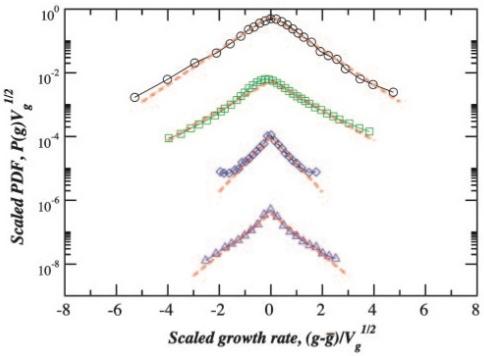
Q: How do complex systems grow ?

Q: How big are the rare events (often neglected by simple models) ?

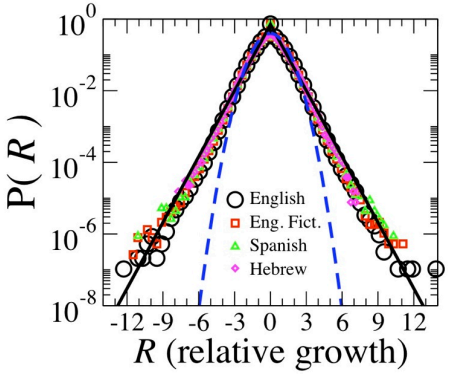
- system size, *S(t)*, at time *t*
- Growth rate R(t) = $g(t) \equiv \log\left(\frac{S(t+1)}{S(t)}\right) = \log S(t+1) - \log S(t)$
- Laplace PDF $P(R) = \frac{1}{\sqrt{2}\,\sigma(R)} e^{-(\sqrt{2}|R-\langle R\rangle|/\sigma(R))}$

Excess number of large growth (+/-) events as compared to the Gibrat multiplicative growth model which predicts a Gaussian distribution for P(R)

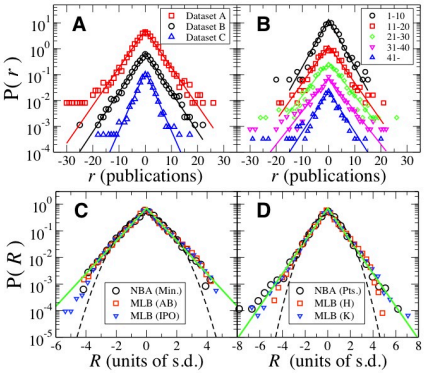
Firm size / Country GDP [1]



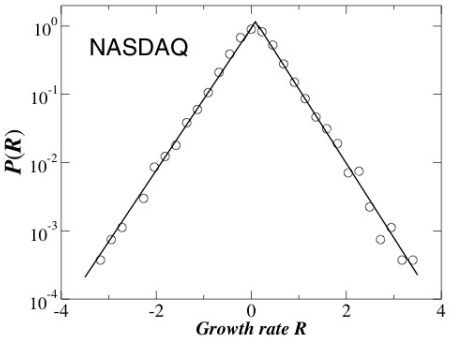

Word Use [2]




Individual Productivity [3]




Stock Price [4]

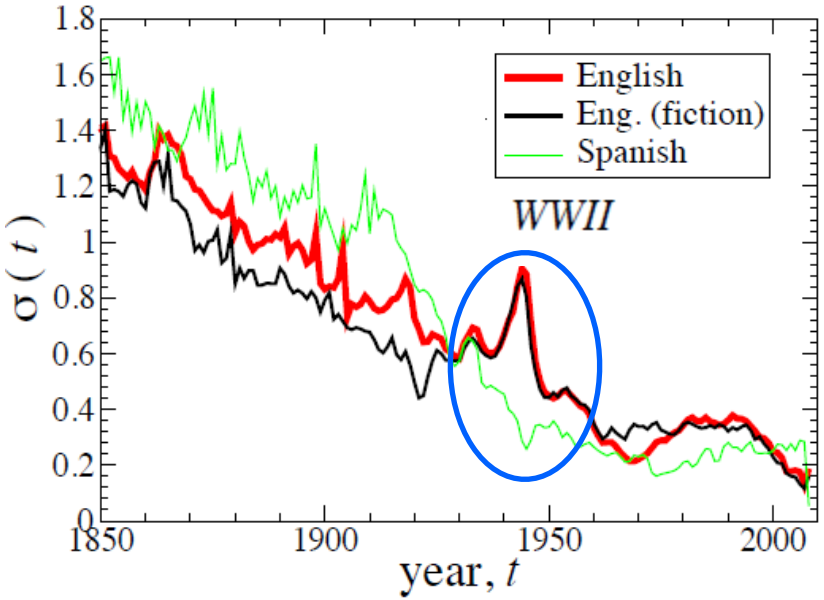




[1] D. Fu, et al., The Growth of Business Firms: Theoretical Framework and Empirical Evidence. Proc. Natl. Acad. Sci. USA 102, 18801 (2005).
[2] A M. Petersen, et al., Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death, submitted.
[3] A M. Petersen, et al., Quantitative Relations between Group Collaboration and the Productivity Growth Dynamics of Individuals, in preparation.
[4] B. Podobnik, et al., Common scaling behavior in finance and macroeconomics. Eur. Phys. J. B 76, 487 (2010).

# Do historical events change the dynamics?

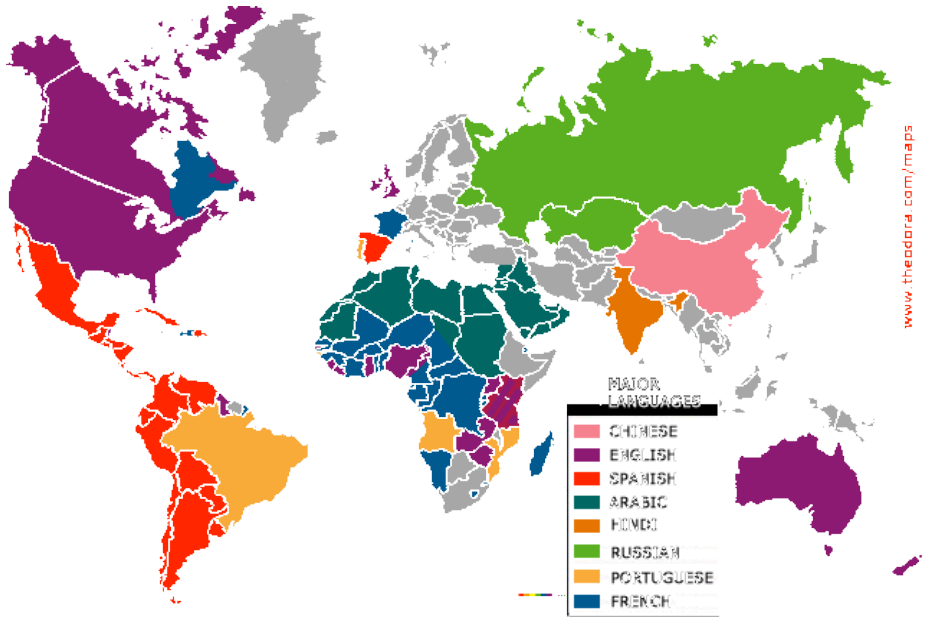
## Spanish speaking countries less involved in WWII




annual growth rates

$$r_i(t) \equiv \ln f_i(t+\Delta t) - \ln f_i(t) = \ln\left(\frac{f_i(t+\Delta t)}{f_i(t)}\right)$$

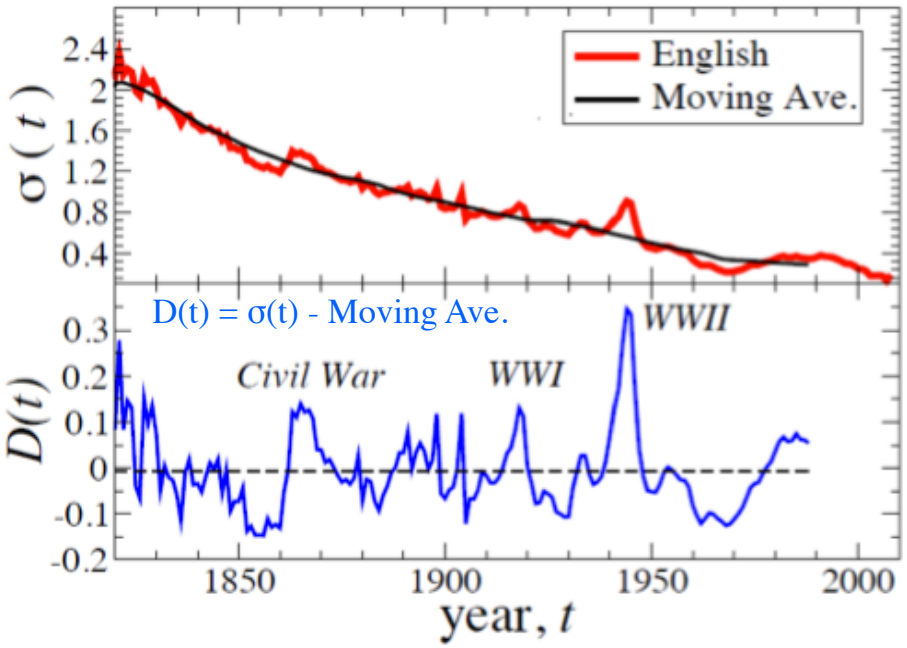
σ(t) = std. deviation of $r_i(t)$




External "shocks" bring more isolated subsystems into contact, leaving outside "ecosystems" (other languages) unperturbed

# Role of political conflict on language



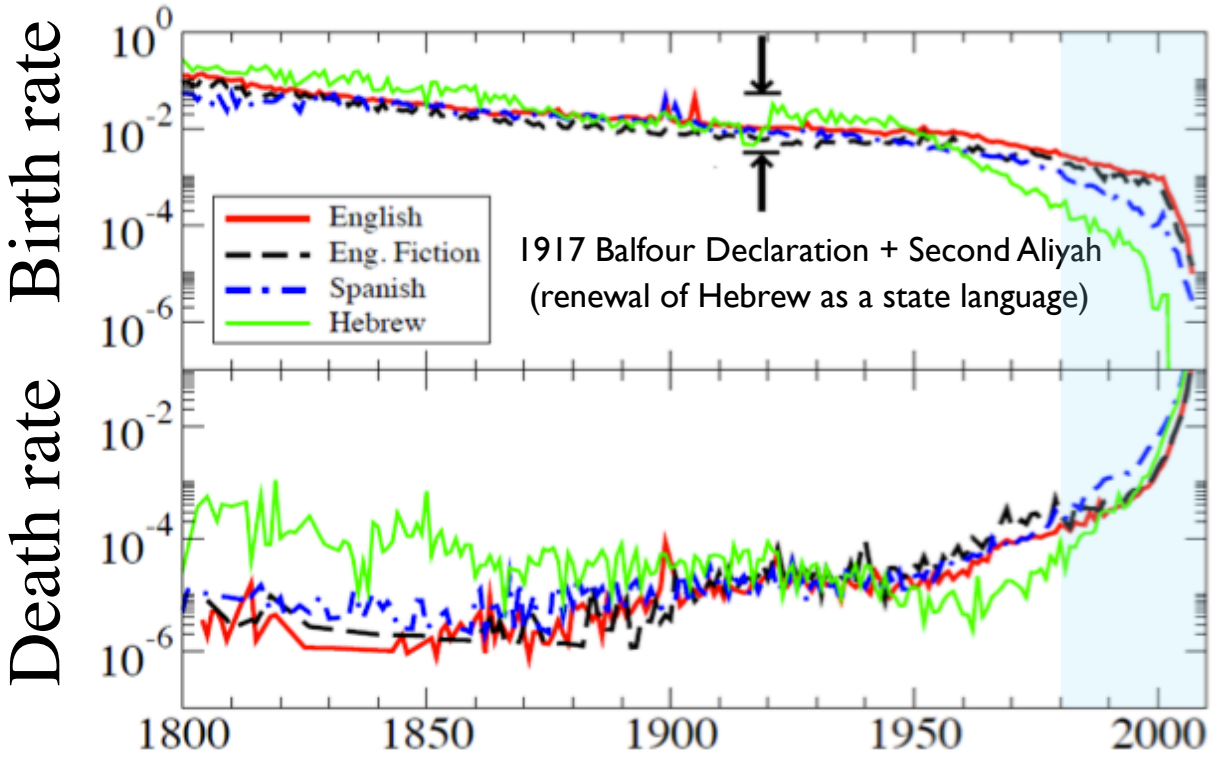




War growth words (peak year)

**Vichyites** (1941)
**Coprosperity** (1942)
**UDSR** (1947)
**fascismo** (1926)
**breechloader** (1940, a type of gun loaded via a magazine instead of through the tip)
**divebomber** (1943)
**Heinkels** (1939) (a type of German bomber)
**sonsabitches** (1944)
**shellshocked** (1944)
**profascist** (1943)
**antifascists** (1945)
**foxtrots** (1946)

Political conflict correspond to periods of increased fluctuations in language, and may serve to increase the rate of cross-fertilization of different languages with new words

# Birth and Death of Words




Era of automatic spell-check editing

The modern era of publishing, which is characterized by more strict editing procedures at publishing houses, computerized word processing (automatic spell-checking) technology, has led to a drastic increase in the death rate of words.

Using visual inspection we verify most changes to the vocabulary in the last 10–20 years are due to the extinction of misspelled words and nonsensical print errors, and to the decreased birth rate of new misspelled variations.

This phenomenon reflects the *decreasing marginal need for new words*. The new words, however, are biased towards words with relatively high frequency.

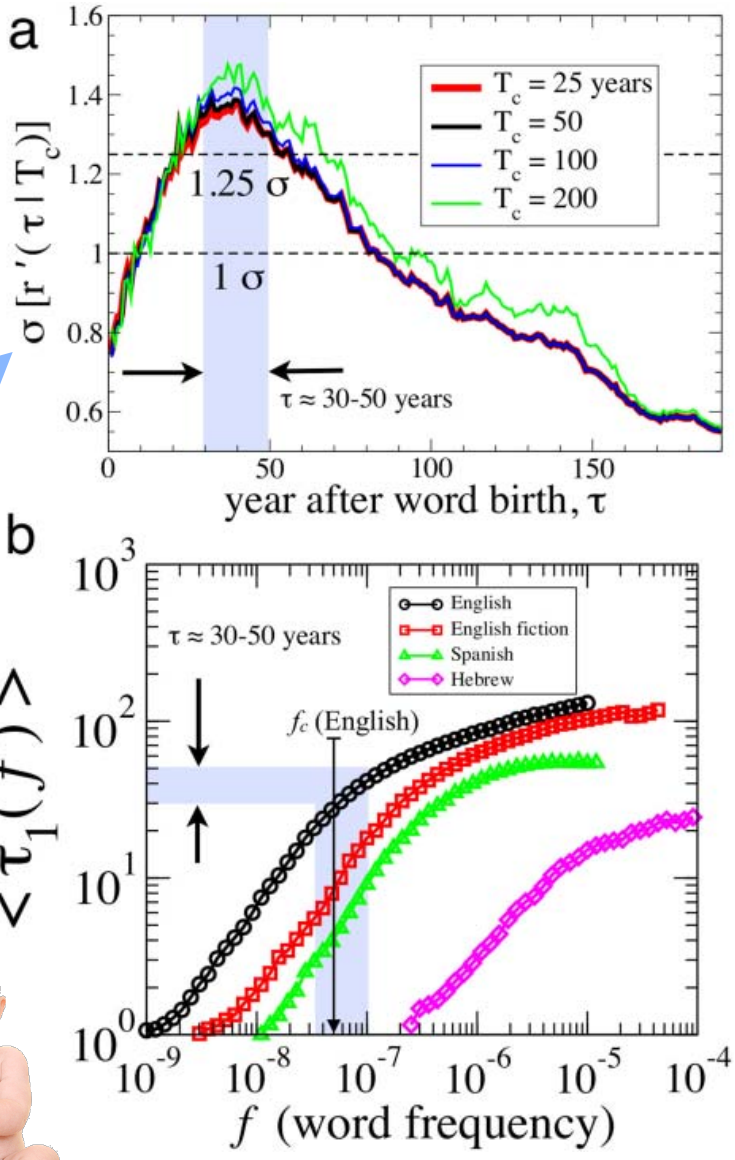
# The life-cycle of a new word

Normalized growth rate of a "new word"

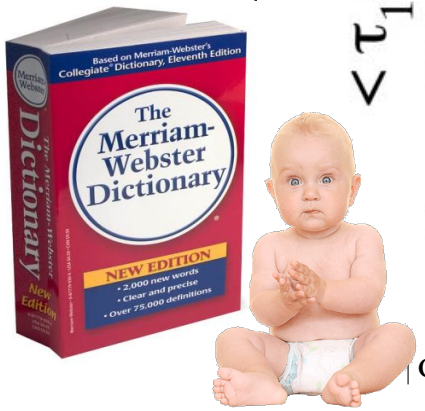
$$r_i'(\tau) \equiv r_i(\tau)/\sigma[r_i]$$

Is there a tipping point in the life-cycle of a new word? The English corpus threshold $fc \equiv 5\times10^{-8}$ maps to the first passage time corresponding to the peak period $t \approx 30 - 50$ years, which is the characteristic generational timescale of humans (and language evolution)

?

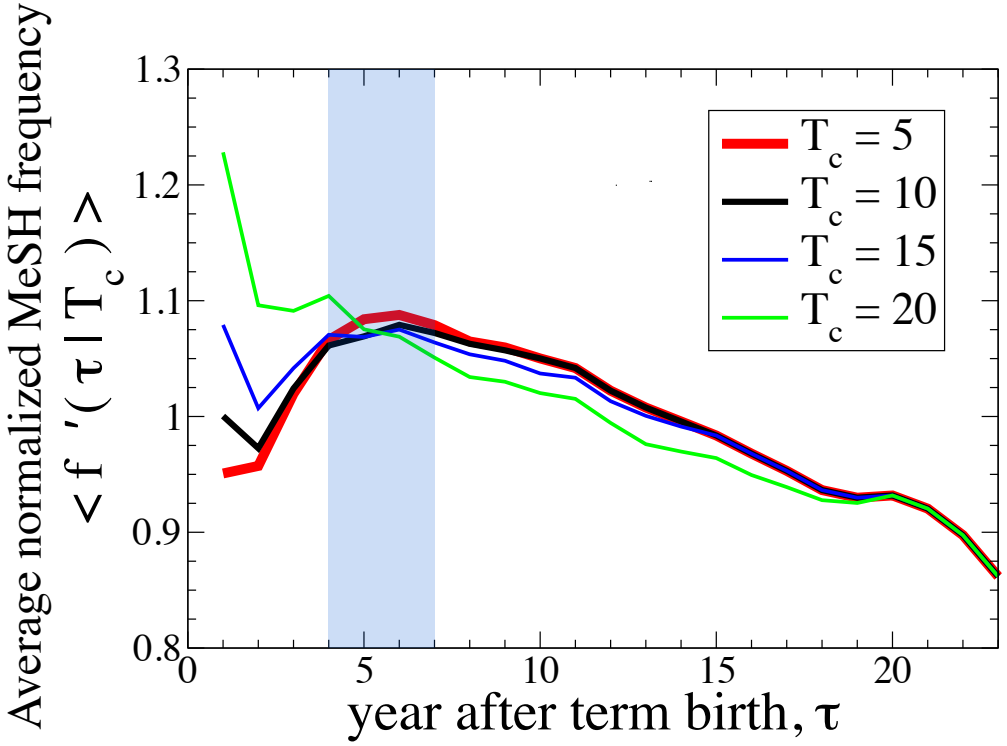



"Both dictionaries had excellent coverage of high-frequency words but less coverage for frequencies below $10^{-6}$ : 67% of words in the $10^{-9}$ to $10^{-8}$ range were listed in neither dictionary" Michel et al., Science (2011)




Quantifying the tipping point for word use.

# Life-cycle analysis of Mesh terms




**The growth trajectory of individual mesh terms.**

Most new MeSH concepts reach their peak popularity around roughly 4-7 years.

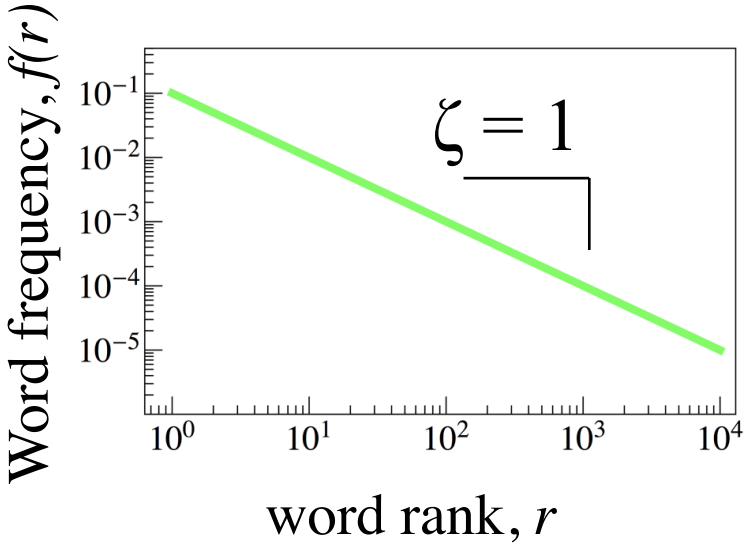
The 4 trajectories are calculated using only MeSH terms with lifetime $L_i > T_c = \{5,10,15, 20\}$ years and birth year $y_i(0) \geq 1987$.

Is there a characteristic life-cycle for scientific trends? 4-7 years is also consistent with the peak in the citation trajectory of highly cited papers

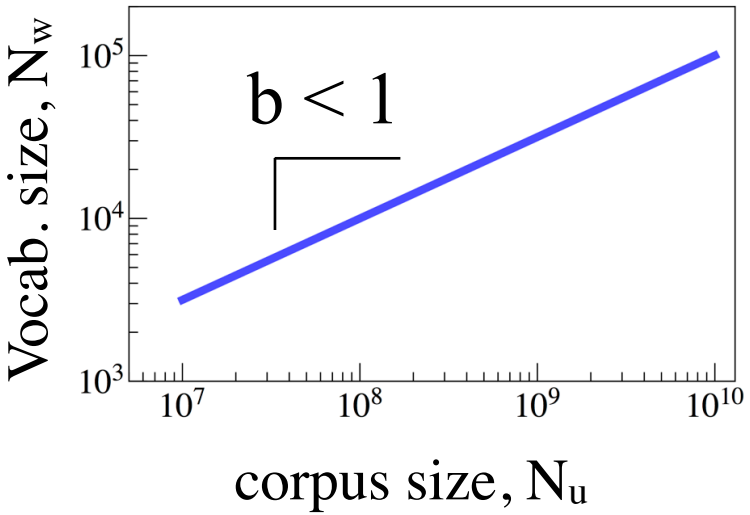
# *Structural evolution of languages across time*

Famous Zipf + Heaps' laws are based on *static* snapshots of (relatively) small texts/corpora

*Zipf's law:* $f(r) \sim 1/r^{\zeta}$



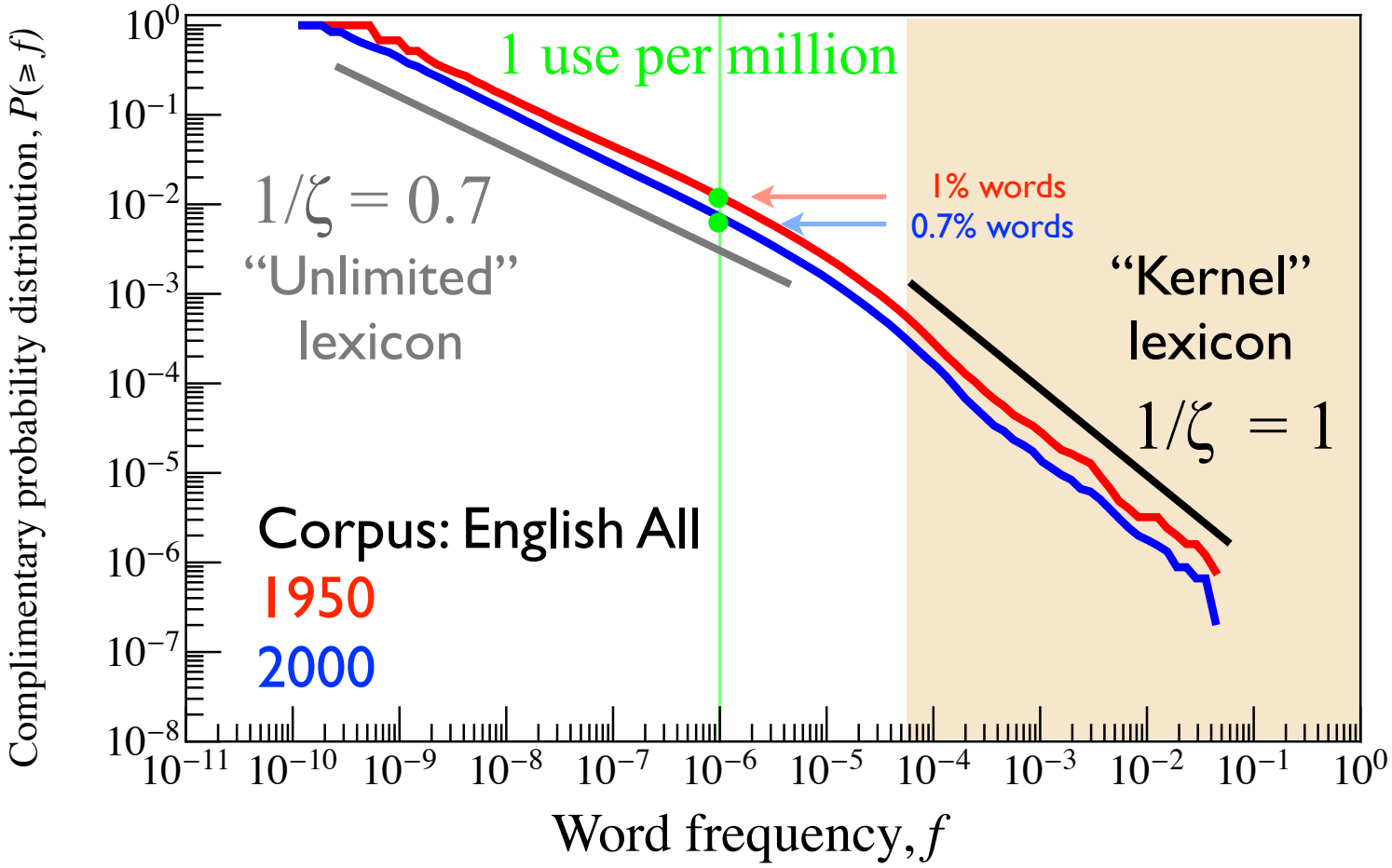

*Heaps' law:* $N_w \sim (N_u)^b$




Q: can we learn anything from analyzing the properties of these statistical laws over time?

# "Dark Language": a hidden Zipf's law

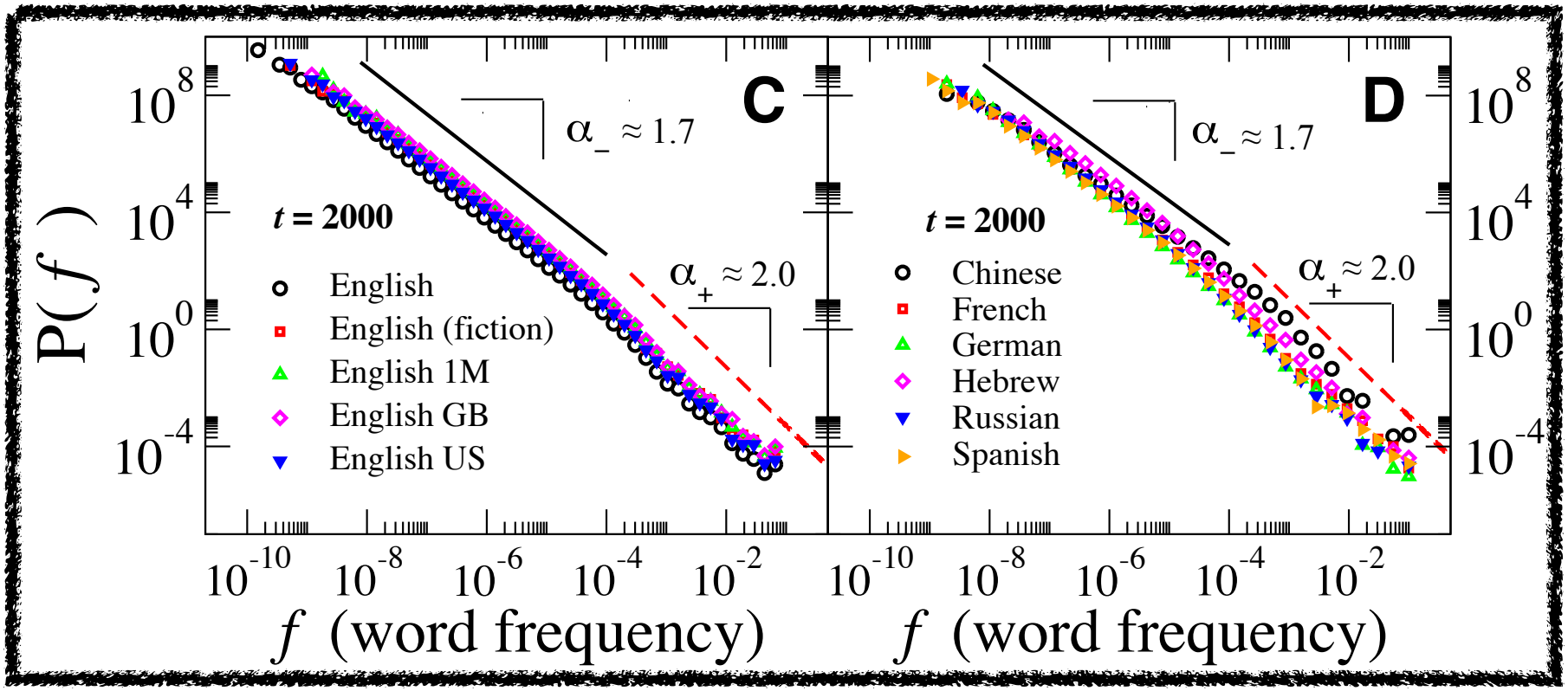
*$P(\geq f)$ is the percentage of 1-grams ("words") with observed frequency larger than $f$*




Hence, dark language* is composed of approximately 99% of the 1-grams recorded in each corpora, leaving only ~1% of words that constitute our "Kernel" lexicon

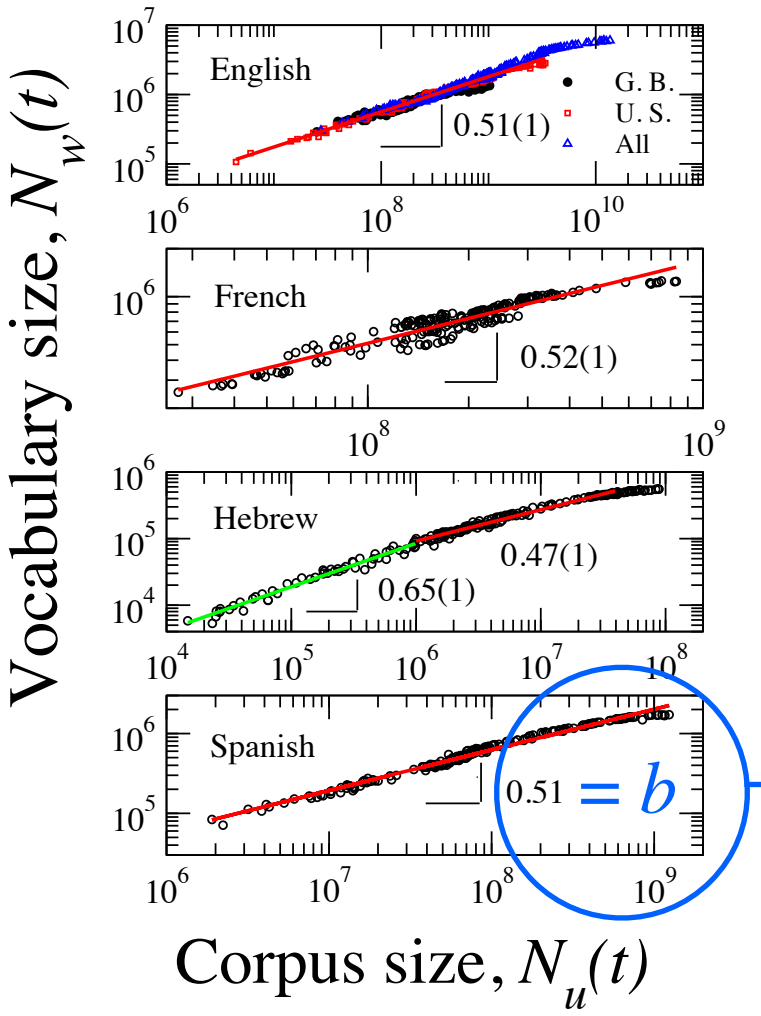
*Recent estimates on the composition of physical matter in the universe: 72.8% dark energy, 22.7% dark matter and 4.6% ordinary matter. Hence, 95% of matter-energy is dark. ("Seven-Year Wilson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results". nasa.gov)

# Consistent patterns of “dark language” across 7 languages




A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley, M. Perc
**Languages cool as they expand: Allometric scaling and the decreasing need for new words**
Scientific Reports 2, 943 (2012)

# Using Heaps' law to reveal the marginal utility of new words




Allometric scaling analysis is used to quantify the role of system size on general phenomena characterizing a system, and has been applied to understand the metabolic (activity) rate of systems with sizes ranging from mitochondria to cities.

Here each data point corresponds to one year: $N_u(t)$ is the total number of "tokens" printed in year $t$ and $N_w(t)$ is the number of distinct tokens in the same year

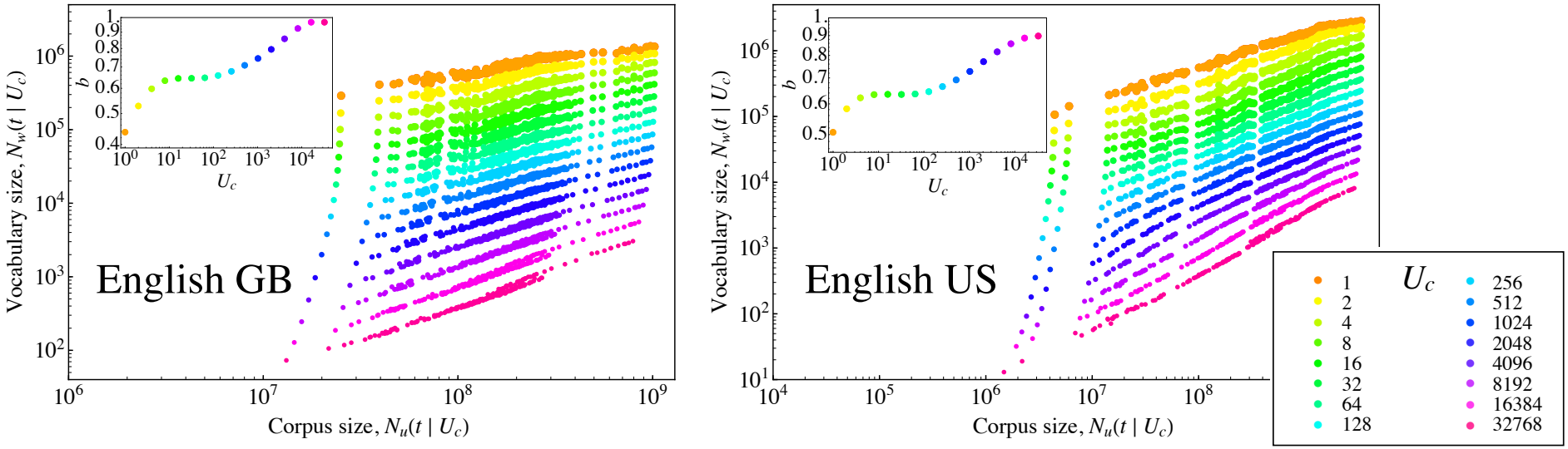
*Heaps' law:* $N_w(t) \sim (N_u(t))^b$

Marginal need for new words (decreasing for $b<1$)

$$\partial N_w / \partial N_u \sim (N_u)^{b-1}$$

$b < 1$ corresponds to an "economies of scale" and implies a decreasing marginal need for additional words as a corpora grows. Because we get more and more "mileage" out of new words in an already large language, additional words are needed less and less. Interestingly, many economic systems have $b > 1$, whereas biological systems have $b < 1$.

# Using Heaps' law to provide insight into the dependency structure between words




Q: How does *b* change if we only include words with $u_i \geq U_C$ in our allometric scaling analysis??

As $U_C$ increases the Heaps scaling exponent increases from $b \approx 0.5$, approaching $b \approx 1$, indicating that core "Kernel" words are structurally integrated into language as a proportional background, $N_u(t) \sim N_w(t)$, quantifying how the kernel lexicon is the structural "glue" with larger marginal utility per word

Institutions Markets Technologies
IMT
INSTITUTE FOR ADVANCED STUDIES LUCCA

# *Food for thought*

- Digitization of historical archives is vastly extending our quantitative perspective on history
- A vast amount o language belongs to an "unlimited" lexicon, consisting of highly specific contextual terminology. Consider that the common everyday words, roughly the top 30,000 most used words which are used with a frequent of more than 1 per million, account for only 1% of the English language vocabulary
- Words compete with irregular forms and synonyms in a competitive environment: "persistence" is gradually suffocating the use of "persistency"
- The growth of language is very sensitive to socio-political shocks, such as war. New words enter largely as a result of technological innovation, but also due to shifts in social behavior: consider that the words "girlfriend" and "boyfriend" emerged only in the early 1960s, likely reflecting a sexual revolution which has major biological implications (e.g. disease spreading, birth rate, etc.). Also, the words "treehuggers" and "ecowarriors" emerged in the early 1990s in conjunction with the "save the earth" movement.
- The sustainability of new and old words likely reflects the word's marginal utility as derived from the implicit dependency structure of language (grammar)

A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley.
**Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death**
Scientific Reports 2, 313 (2012).

A. M. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley, M. Perc
**Languages cool as they expand: Allometric scaling and the decreasing need for new words**
Scientific Reports 2, 943 (2012)

Thank You!

A special thanks to my collaborators:
Joel Tenenbaum, Matjaz Perc,
Shlomo Havlin, Gene Stanley

*http://physics.bu.edu/~amp17/*

Title: Using big data to quantify the evolution of written corpora at the micro and macro scale

Abstract:

Generic evolutionary forces of survival and reproduction are believed to drive the evolution of language. Using the Google Inc. n-gram dataset spanning 200+ years, we show patterns consistent with competitive dynamics at the level of individual words (tokens) as well as at the level of entire corpora. At the micro scale, we demonstrate tipping points in the life-cycle of new words, growth patterns consistent with competition for limited “market opportunities”, and evolutionary selection induced by modern editing software (Petersen et al, Sci. Reports 2012). At the macro scale we show that languages “cool as they expand”, a dynamic property that highlights periods of political conflict which are characterized by heightened levels of language fluctuations (Petersen et al, Sci. Reports 2013). We will show that these general methods can be extended to other evolving categorical systems such as the MeSH (Medical Subject Headings) vocabulary used by the United States National Library of Medicine.