Using big data to quantify the evolution of language at the micro and macro scale

Alexander M. Petersen

IMT Lucca Institute for Advanced Studies, Lucca 55100, Italy
• “Digital Humanities” and “Culturomics”: new science made possible by “crowd-sourced” “Big data”

• Google digital books: 5 million books and 500 billion word uses

• Competition (for limited use, attention)

• Geographic variation: the role of socio-political shocks

• Tipping points in the life-cycle of new words

• Languages become “colder as they expand”

• Uncovering an enormous hidden “Dark language”
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 billions of word counts from books
https://books.google.com/ngrams

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

Language is a structured system

Association network of ~ 4000 anatomical words
A word network constructed from ~20,000 biomedical terms (MeSH: medical subject headings) developed by the
US National Library of Medicine

[A]  Anatomy
[B]  Organisms
[C]  Diseases
[D]  Chemicals and Drugs
[E]  Analytical, Diagnostic and Therapeutic Techniques and Equipment
[G]  Biological Sciences
Language is a competitive system

Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death
Scientific Reports 2, 313 (2012).
Evidence for competition in a limited marketplace

Is this a:

a) Xray
b) Radiogram
c) Roentgenogram

??
**Words compete for limited market share**

\[ u_i(t) : \text{# of uses of word } i \text{ in year } t \]

\[ f_i(t) \equiv u_i(t)/N_u(t), \]

**Competition between:**

- Synonyms
- Spellings (e.g. color vs. colour)
- other ideas in an abstract “idea space”. Consider the Euphemism treadmill:

  shell shock (WWI) \(\Rightarrow\)

  battle fatigue (WWII) \(\Rightarrow\)

  operational exhaustion (Korean War) \(\Rightarrow\)

  PTSD (Vietnam War)
Competition in subtle spelling variations

“Rich get richer” and the survival of the fittest....
Geographic variation in the battle of the (ir)regular verb conjugations: the past tense “-ed”, “-t”, ....

Digital traces of cultural Nostalgia & Optimism

How often do we dream about the future? and
How often do we refer to the past?

Quantitatively, analysis of culture using millions of digitized books shows that the value of this future orientation index for 45 countries describes how much Internet users seek information about the future rather than the past. The analysis quantifies to what extent Internet users worldwide are seeking information about the previous year (in red) and the following year (in blue). Data are plotted for each of 52 weeks per year. The strong tendency for countries in which there is a focus on the future supports economic success and the relationship between the economic success of a country and the type of information sought online, perhaps due to economic influences on available Internet infrastructure. We suggest that the second finding may reflect international differences in the orientation index.

The frequency with which the word “2009” appeared in search engine query data increased sharply in 2008, declined gradually until 2010, and then increased again in 2011. The median frequency of the searches from year to year shows a strong correlation with per capita GDP in 2010 for 45 countries.

Collective memory has both a short-term and a long-term component. Observing the media trajectory of the 25 most famous personalities born between 1865 and 1920, the 1865 cohort characterized by four parameters: (i) initial age of celebrity (70 years after birth, tick mark), and (ii) age of peak celebrity (4 years, blue line); (iii) time of the subsequent rise to fame (4 years, blue line); (iv) half-life of the post-peak forgetting (73 years, red line). Inset: The media trajectory of the 25 most famous personalities born between 1883 and 1950 shows that the media trajectory is getting shorter and the media trajectory of the 25 most famous personalities born between 1883 and 1950 shows a regime marked by slower forgetting: more rapid assimilation of the new (blue), and our past faster with each passing year (Fig. 3A).

Quantifying the advantage of looking forward. Preis et al. (2013) Scientific Reports.

Author contributions: T.P., H.S.M., H.E.S. and S.R.B. performed analyses, discussed the results, and contributed to the preparation of the manuscript. J.G. created and compiled search engine query data, and created the graphics. T.P. and H.S.M. wrote the manuscript. All authors reviewed the manuscript.

Competing financial interests: The authors declare no competing financial interests.

This work was partially supported by the German Research Foundation Grant PR 1305/1-1

http://creativecommons.org/licenses/by-nc-sa/3.0/
Median frequency (log) and Frequency vs. Year. The trends show how cultural turnover is accelerating. We forget faster. The half-life of the post-peak forgetting phase (73 years, red line). Inset: The media trajectory of the 25 most famous personalities born between 1800 and 1920 in various careers. (i) age of peak celebrity (70 years after birth, tick mark), and (ii) doubling time and half-life over time. (iii) age of peak celebrity (4 years, blue line); (iv) age of peak celebrity (34 years old, tick mark); (v) half-life of the post-peak forgetting phase (73 years, red line). Inset: We forget faster. The half-life of the post-peak forgetting phase (73 years, red line). Inset: We forget faster. The half-life of the post-peak forgetting phase (73 years, red line). Inset: We forget faster. The half-life of the post-peak forgetting phase (73 years, red line).
Historical gastronomy. (most famous of the groups (1 other states), they rapidly rose to become the when, upon being elected president of the United
demiology: frequency of 2 rises slowly thereafter. (Their fame peaked at a
around 30. But the fame of the actors we studied, 
fame grew as a function of age (Fig. 3F and fig. S10).
tors, artists, writers, politicians, biologists, phys-
tics, and mathematicians), examining how their
factors, the challenge of culturomics lies in the in-
Writers become popular with a frequency 5 10−6 at some point. (In contrast, from 1946 to 1954,
incubated in the 1920s and championed in the 1960s
feminism (English)
féminisme (French)

Let’s talk about SEX

Changing norms of sexual equality in our society

Quantitative Analysis of Culture Using Millions of
With its roots in the first perceived sexual revolution in the 1920s, this 'revolution' in 1960s America encompassed many groups who are now synonymous with the era. Feminists, gay rights campaigners, hippies and many other political movements were all important components and facilitators of change.
Ok Let’s **Really** talk about SEX

![Graph showing the word frequency of "friends with benefits" and "fuck buddy" over time](image)

Evolution of not only terminology representing social norms....
Ok Let’s **Really** talk about SEX

![Google books Ngram Viewer](image)

but cultural evolution of sexual norms also has significant implications for disease control and human reproduction...
Do historical events change the dynamics?

Spanish speaking countries less involved in WWII

Annual growth rates

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

\[ \sigma(t) = \text{std. deviation of } r_i(t) \]
Political conflict causes periods of increased fluctuations in language and an increased rate of cross-fertilization between languages.
Languages “cool as they expand”

annual growth rates

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

\[ \sigma(t) = \text{std. deviation of } r_i(t) \]

\[ \sigma(t) = \text{std. deviation of } r_i(t) \]

measures the characteristic fluctuations in word growth

\[ \sim \text{“system temperature”} \]

Q: Is language evolution slowing down?

Q: What is the counteractive role of new language platforms?

e.g. text messaging, Twitter

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

The birth rate has also decreased, indicating the decreasing marginal need for new words. However, the new words that do survive have relatively high word use frequency (intrinsic fitness, e.g. e-mail, Google).
The life-cycle of a new word

Is there a tipping point in the life-cycle of a new word?

New words demonstrate peak “instability” when they are $\approx 30 - 50$ years old, corresponding to:

- a) the typical time it takes to be accepted into a dictionary
- b) the generational timescale of humans (and language evolution)

![Graph showing fluctuations in word use over time, with peaks around 30-50 years]
“Dark Language”: a hidden Zipf’s law

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

Only $\sim 1\%$ of words in each corpora belong to the "Kernel" lexicon (words that a typical person could recognize)

A vast hidden "Dark language" (Unlimited Lexicon) accounts for approximately 99% of the 1-grams recorded in each corpora,
*Recent estimates indicate that 95% of the universe is composed of dark matter/energy (72.8% dark energy, 22.7% dark matter), and only the remaining 4.6% ordinary matter.

("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

\[
P(f) \sim f^{-\alpha}\]

\begin{align*}
t &= 2000 \\
\alpha_+ &\approx 2.0 \\
\alpha_- &\approx 1.7
\end{align*}

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)
Food for thought

- Digitization of historical archives is vastly extending our quantitative perspective on history.

- A vast amount of 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).

Thank You!

A special thanks to my collaborators:

Joel Tenenbaum, Matjaz Perc,
Shlomo Havlin, Gene Stanley

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

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

Abstract:

What if you could analyze every word every book in every library, bookcase, and attic around the world? What kind of trends and changes in culture could you discover? All of the world's books constitute an immense “crowd-sourced” historical record that traces the evolution of culture beyond the limits of oral history. But to analyze individual words over time has been incredibly painstaking-- until now. Google has digitized a huge collection of written language in the form of the Google Books Ngram Viewer web application (https://books.google.com/ngrams). 4% of all books ever published have been digitally scanned, making 10 million histories for individual words, a vast archive of cultural dynamics over more than two centuries. With statistical methods borrowed from physics, we show what the frequencies of words can tell us about every aspect of society, from the recent emergence of the environmentalism to the impact of feminism on human sexual behavior over the last 200+ years, from the the impact of globalization on vocabularies in 7 languages, to the role of spell-checkers on the survival rate of "mutant" words.
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 $b<1$)

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

$b < 1$ 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
Life-cycle analysis of Mesh terms

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?
"zero sum" competitive system

\[ P(r') \equiv \ln f_i(t+\Delta t) - \ln f_i(t) = \ln \left( \frac{f_i(t+\Delta t)}{f_i(t)} \right) \]

\[ r = \text{annual growth rates in the word usage frequency} \]

\[ r_i(t) = \text{first instance of} \]

**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: \( f_c = 10^{-5} \)

\( P(r) \) is centered around \( r \approx 0 \), 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)?

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]