DECODING SOCIETY:

HOW TO USE NATURAL LANGUAGE PROCESSING TO UNDERSTAND BEHAVIOR AT SCALE

Thomas Hills
University of Warwick
for **HERSS Summer School, September, 2024**

WARWICK
THE UNIVERSITY OF WARWICK

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

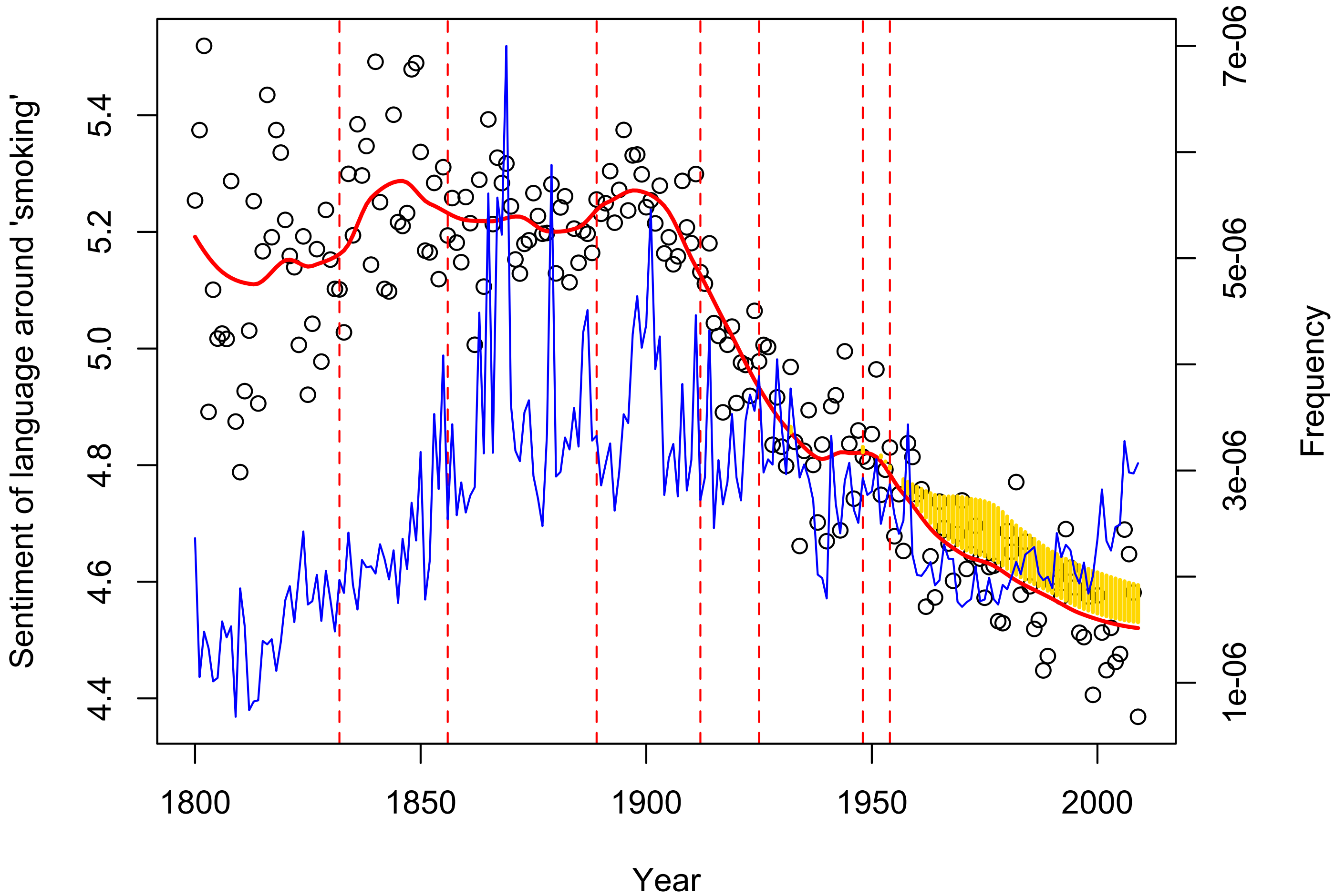
The Alan Turing Institute

THE ROYAL SOCIETY

The Leverhulme Trust

BRITISH ACADEMY
for the humanities and social sciences

The history of the word ‘smoking’



What is Natural Language Processing?

Counting and advanced counting

- Natural language processing is a “distant read”

  Computational method of content analysis, allowing one to ‘read’ millions of words of text with highly reproducible results.

  Effectively what we have the computer do is count things (frequencies and relationships)

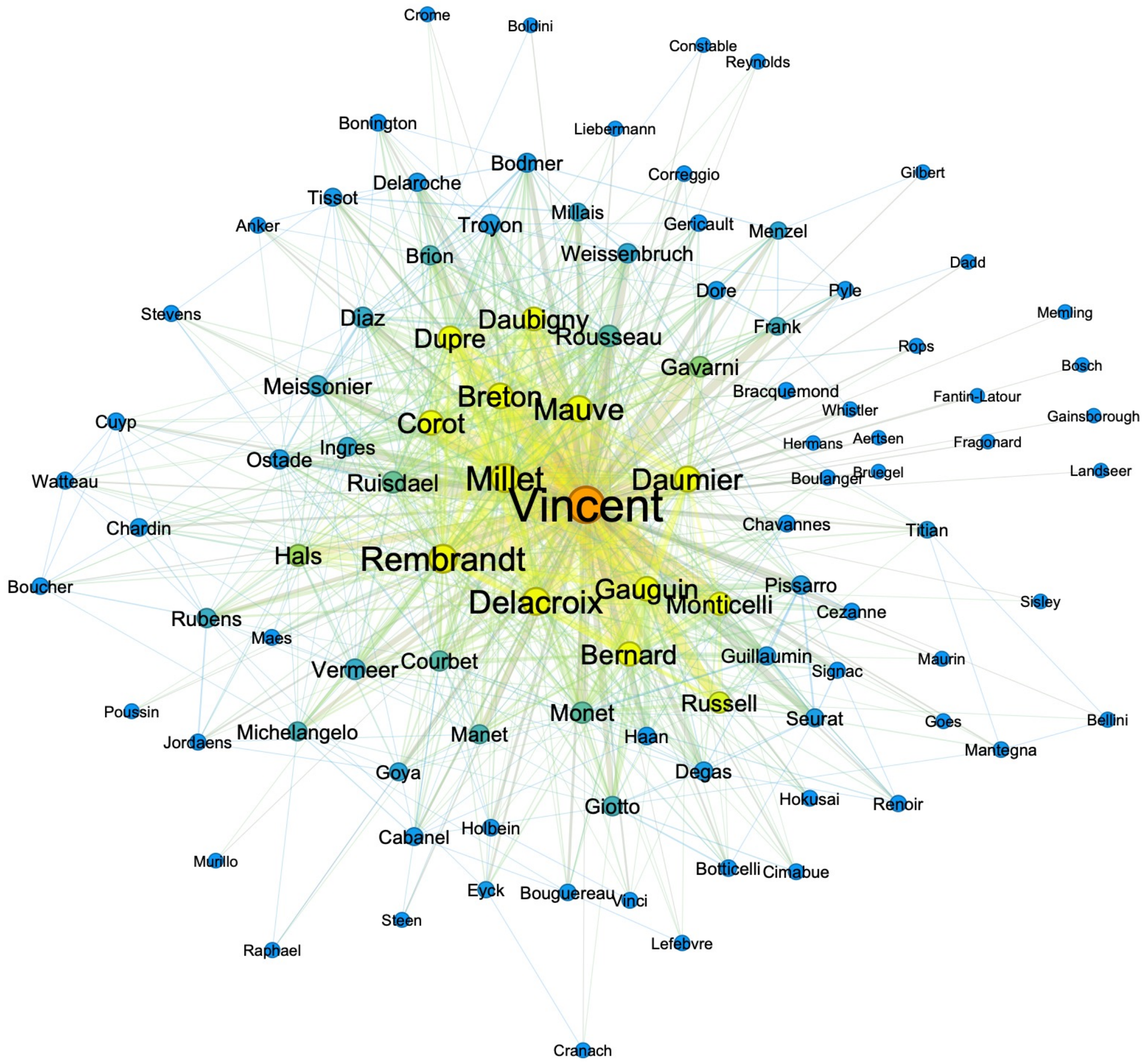
  Highly reproducible because you can and should share the code and the data

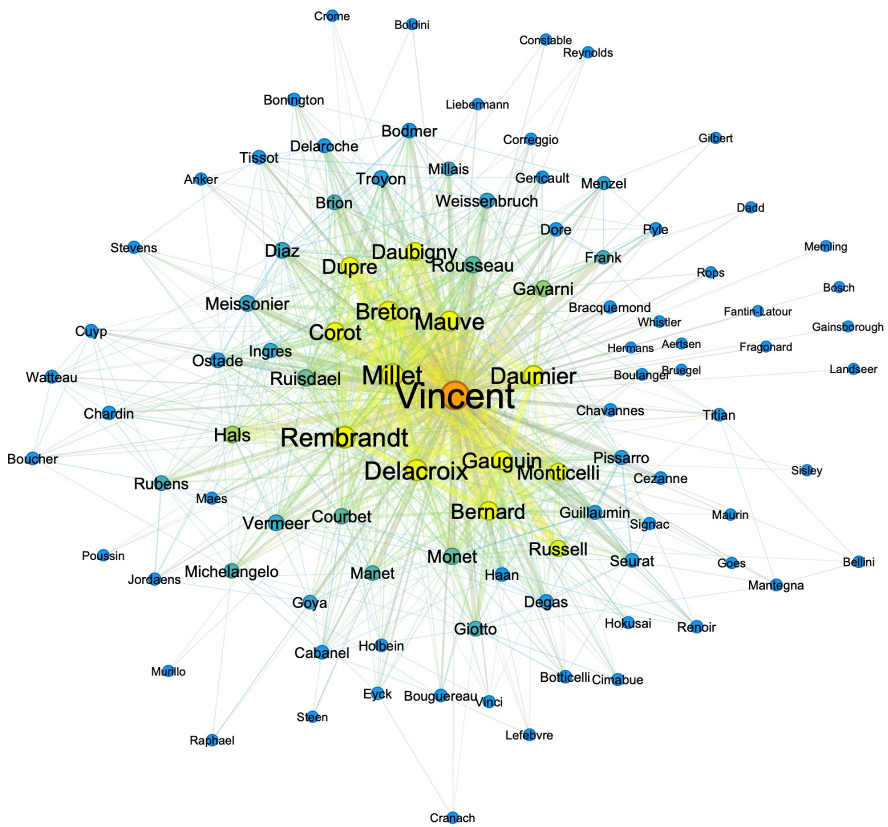
Two problems in NLP:

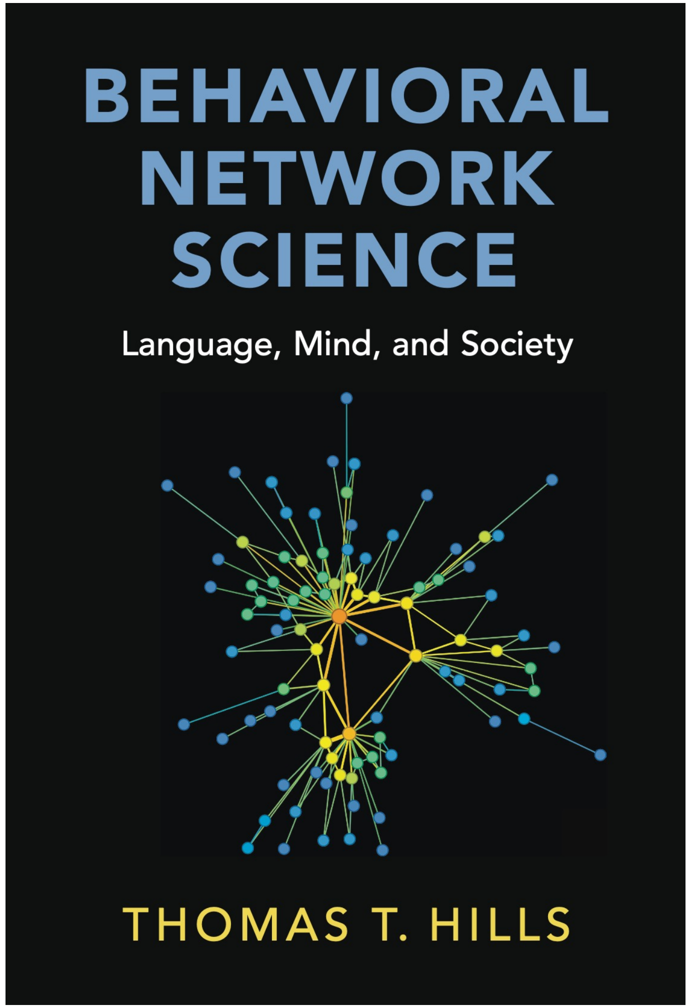
Big problem: How do we ask good questions and do good research when we’re thinking about language and NLP?

Small problem: What are the methods?

Van Gogh's Influencers




Crome
Boldini
Constable
Reynolds
Bonington
Liebermann
Bodmer
Delaroche
Tissot
Correggio
Gilbert
Anker
Troyon
Millais
Gericault
Menzel
Brion
Weissenbruch
Dadd
Dore
Pyle
Stevens
Diaz
Daubigny
Dupre
Rousseau
Frank
Memling
Rops
Gavarni
Bosch
Meissonier
Breton
Mauve
Bracquemond
Fantin-Latour
Cuyp
Corot
Whistler
Gainsborough
Ingres
Hermans
Aertsen
Fragonard
Ostade
Landseer
Watteau
Ruisdael
Millet
Daumier
Boulanger
Bruegel
Vincent
Chardin
Chavannes
Titian
Hals
Rembrandt
Boucher
Pissarro
Gauguin
Sisley
Delacroix
Monticelli
Cezanne
Rubens
Maes
Bernard
Guillaumin
Maurin
Courbet
Vermeer
Signac
Russell
Poussin
Monet
Seurat
Goes
Bellini
Michelangelo
Jordaens
Manet
Haan
Mantegna
Goya
Degas
Giotto
Hokusai
Renoir
Cabanel
Holbein
Murillo
Botticelli
Cimabue
Eyck
Bouguereau
Vinci
Steen
Lefebvre
Raphael
Cranach


BEHAVIORAL
NETWORK
SCIENCE
Language, Mind, and Society
THOMAS T. HILLS

High level view of NLP

- Two approaches:
  1. We *quantify* some feature of the text (how positive or negative, how concrete/abstract, how complicated, how ‘truthful’, etc.)
  2. We *categorize* the data in some way (e.g., What’s the topic? Is it spam?)

Two things we need to do good NLP?

- **Operational Definitions:** clear definitions that allow us to count things (frequencies and relationships)
  - Based on analytical constructs (i.e., theory)
- **Language data:** What is the data quality? Is it representative? How reliable is it? Do we have enough?

Main approaches in NLP

1. Counting words
2. Word feature analysis (sentiment)
3. Word and document similarity
4. Topic modelling

1.

Counting words/ Dictionary methods

What words matter?

Question: Is society becoming more selfish/individualistic?

Word counting

- Beautiful story about historical psychology based on single word analyses from an off-the-shelf resource: Google Ngram Viewer

Psychological Science
24(9) 1722–1731

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613479387
pss.sagepub.com
SAGE

The Changing Psychology of Culture From 1800 Through 2000

Patricia M. Greenfield
Department of Psychology, University of California, Los Angeles

Abstract

The Google Books Ngram Viewer allows researchers to quantify culture across centuries by searching millions of books. This tool was used to test theory-based predictions about implications of an urbanizing population for the psychology of culture. Adaptation to rural environments prioritizes social obligation and duty, giving to other people, social belonging, religion in everyday life, authority relations, and physical activity. Adaptation to urban environments requires more individualistic and materialistic values; such adaptation prioritizes choice, personal possessions, and child-centered socialization in order to foster the development of psychological mindedness and the unique self. The Google Ngram Viewer generated relative frequencies of words indexing these values from the years 1800 to 2000 in American English books. As urban populations increased and rural populations declined, word frequencies moved in the predicted directions. Books published in the United Kingdom replicated this pattern. The analysis established long-term relationships between ecological change and cultural change, as predicted by the theory of social change and human development (Greenfield, 2009).

Keywords

sociocultural factors, values, cultural change, content analysis, quantitative analysis

Received 8/9/12; Revision accepted 1/16/13

Word counting

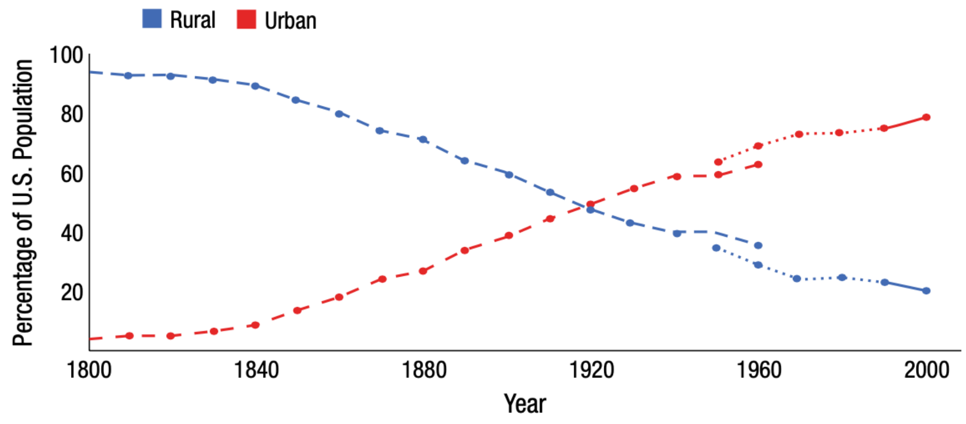


Fig. 1. Percentage of U.S. population living in rural and urban areas from the years 1800 to 2000. Data were drawn from the following sources—1800–1980: U.S. Census Bureau (2004); 1990: U.S. Census Bureau (1992); and 2000: U.S. Census Bureau (2004). The definition of *urban population* changed over the years, and two different definitions were both used in 1950 and 1960, so there are double data points for those years (for details, see Ecological Analysis in the text).

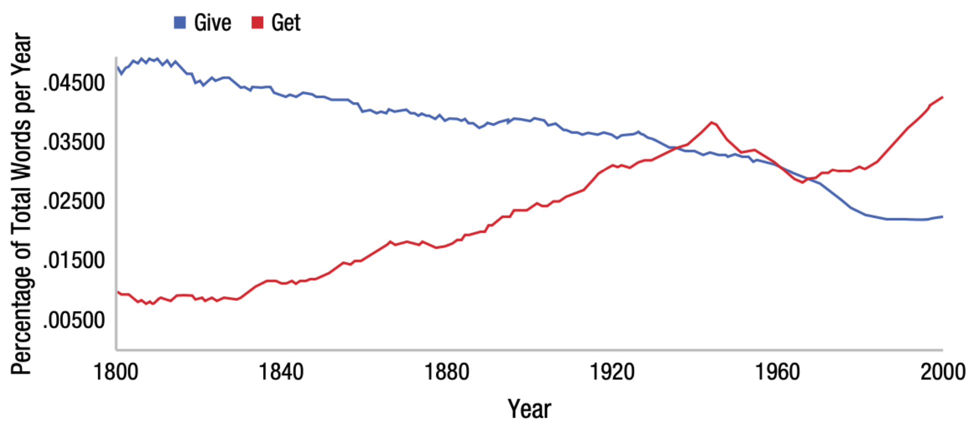
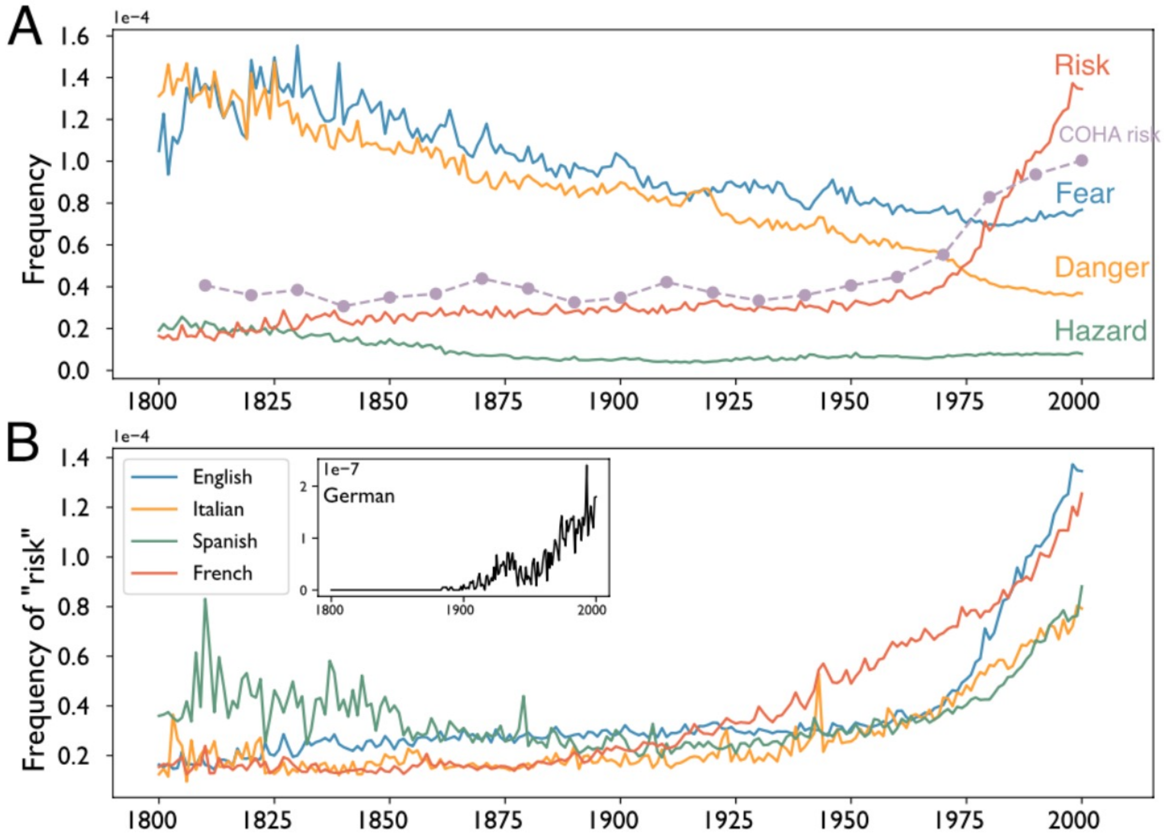


Fig. 3. Frequency of the words "give" and "get" in the Google corpus of American English books from the years 1800 to 2000. The graph was made with the Google Books Ngram Viewer (Michel et al., 2011), with a smoothing of 3.

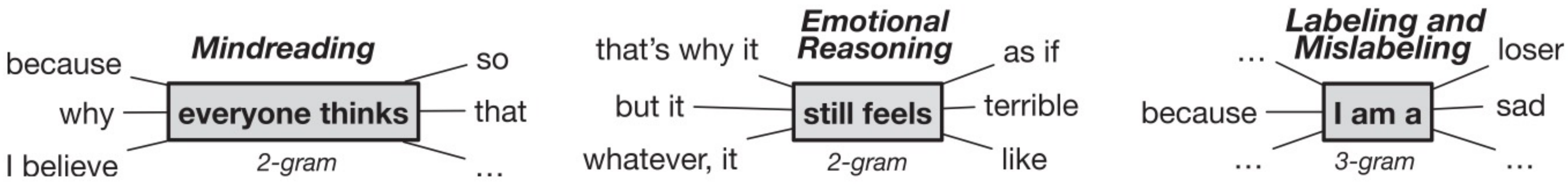
Question: Are we increasingly living in a risk focused society?

“A brief history of risk”



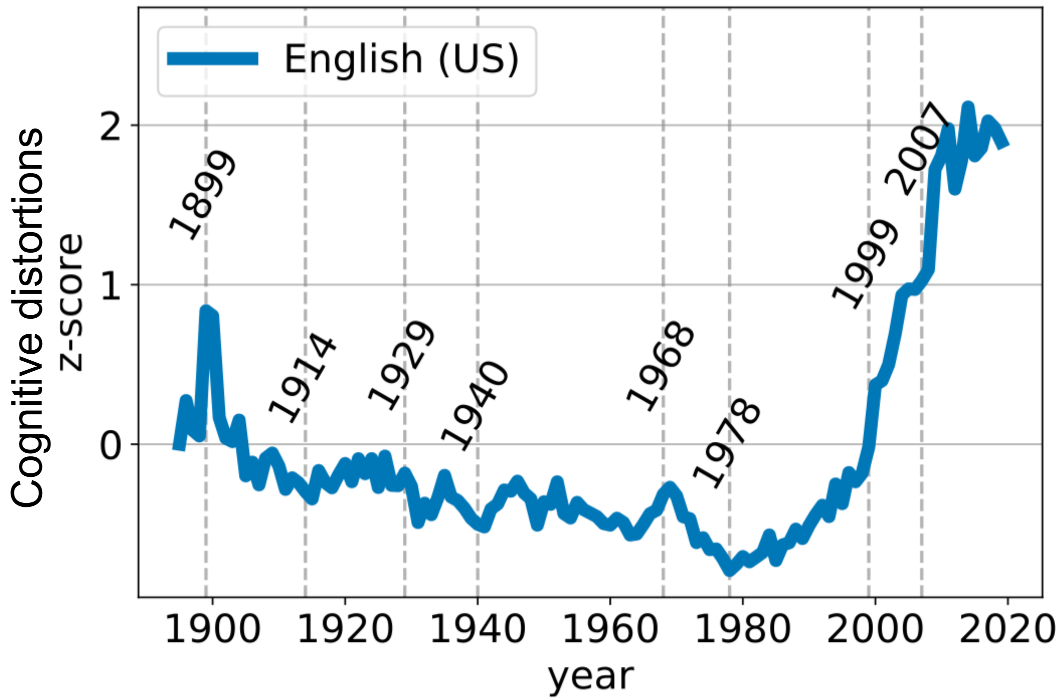
Li, Hills, & Hertwig, 2020

Question: Are our thought patterns becoming increasing maladaptive (e.g., overly negative)?



A dictionary of Cognitive Distortions

Bollen et al. developed a dictionary of cognitive distortions designed by a team of Cognitive Behavioural Therapy experts.



Historical language records reveal a surge of cognitive distortions in recent decades

Johan Bollen[a,1], Marijn ten Thij[a,b], Fritz Breithaupt[c], Alexander T. J. Barron[a], Lauren A. Rutter[d], Lorenzo Lorenzo-Luaces[d], and Marten Scheffer[e]

Word counting: Basic Tools to get started

- Tools:
  - **Google ngram viewer**
  - **LIWC:** off-the-shelf package with built in dictionaries
  - **Google trends**

2.

Word features

What do words represent?

Question: Has American English become more or less concrete over the last 200 years?

Concrete vs. abstract: Concrete means its easier to visualize or see in your mind’s eye

- Children have always a sympathy in the agitations of those connected with them; always, especially, a sense of any trouble or impending revolution, of whatever kind, in domestic circumstances; and therefore Pearl, who was the gem on her mother's unquiet bosom, betrayed, by the very dance of her spirits, the emotions which none could detect in the marble passiveness of Hester's brow. (Nathanial Hawthorne, *The Scarlet Letter,* 1850).

- If you're looking for sympathy you'll find it between shit and syphilis in the dictionary. (David Sidaris, *Barrel Fever,* 1994)

- From *Nature*
- When so little is really known about evolution, even in the sphere of organic matter, where this grand principle was first prominently brought before our notice, it may perhaps seem premature to pursue its action further back in the history of the universe. (Blanshard, 1873)
- Each sex is part of the environment of the other sex. This may lead to perpetual coevolution between the sexes, when adaptation by one sex reduces fitness of the other. (Rice, 1996)

Is American English becoming more concrete over the last 200 years?

- Concreteness norms:
- On a 5-point scale: How concrete is *China*? How concrete is *essentially*.
- Concrete words are recognised faster, more easily recalled; more interesting, more truthful, easy to understand (Brysbaert et al., 2014)
- We have concreteness norms for 40,000 words from Brysbaert et al, 2013.

How to measure document level concreteness

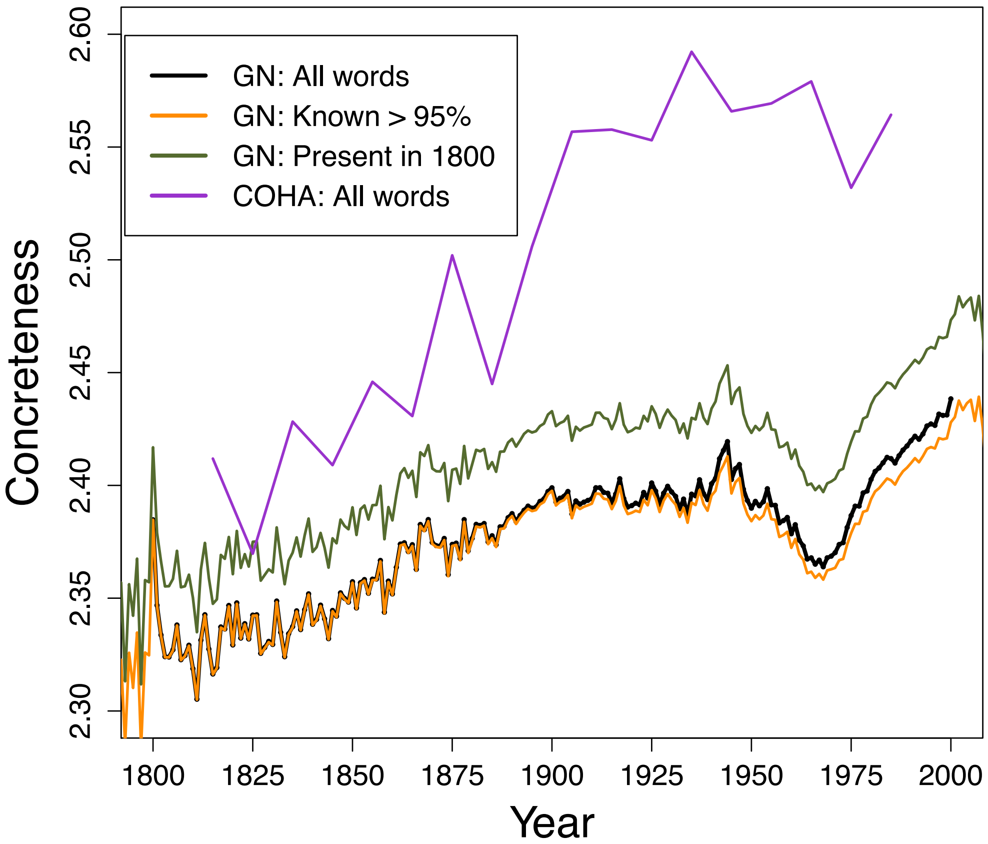
- We take the average concreteness of all words in the norms for each document (year)
- Weighted average
- c_i concreteness score
- p_i = proportion

	value	count
dog	4	1
animal	5	1
organ	3	1
wolf	2	2

$$C_y = \sum_{i}^{n} c_i p_{i,y}$$

Concreteness is rising in American English

- Unit of analysis is ‘year’
- Average concreteness computed per year
- Two different corpora (independent samples)
- This is happening within nouns/verbs, articles.
- Once you have a result you can show it to everyone and collect criticism! New hypotheses!



List of word norms

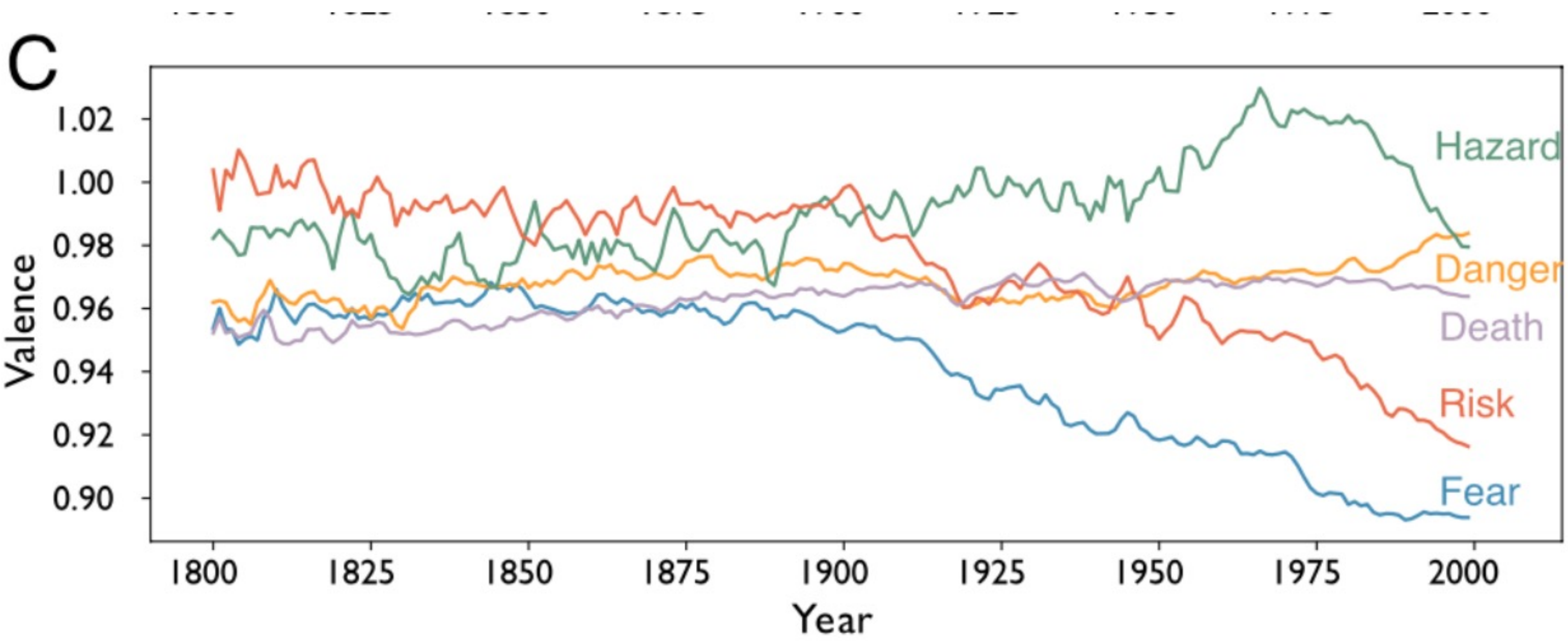
- Valence (sentiment analysis): positive v. negative
- Concreteness: (concrete v. abstract)
- Age of acquisition: at what age learned?
- Free associations: what’s it related to?
- Humor: how funny is it?
- Many more (see Hills & Miani’s ‘Short Primer on Historical Natural Language Processing’—the text for this presentation)

Question: Is ‘risk’ becoming a more negative word in modern society?

Windowing

- We can define words in a ‘window’ around a word of interest, and treat these as ‘documents’

Windowing example: Risk is becoming more negative



Question: How can we tell how happy people were in the past?

Sentiment Analysis: Historical Estimates of Subjective Wellbeing from Millions of Digitized Books

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0750-z

Historical analysis of national subjective wellbeing using millions of digitized books

Thomas T. Hills[1,2*], Eugenio Proto[3,4,6], Daniel Sgroi[3,5] and Chanuki Illushka Seresinhe[2]

In addition to improving quality of life, higher subjective wellbeing leads to fewer health problems and higher productivity, making subjective wellbeing a focal issue among researchers and governments. However, it is difficult to estimate how happy people were during previous centuries. Here we show that a method based on the quantitative analysis of natural language published over the past 200 years captures reliable patterns in historical subjective wellbeing. Using sentiment analysis on the basis of psychological valence norms, we compute a national valence index for the United Kingdom, the United States, Germany and Italy, indicating relative happiness in response to national and international wars and in comparison to historical trends in longevity and gross domestic product. We validate our method using Eurobarometer survey data from the 1970s and demonstrate robustness using words with stable historical meanings, diverse corpora (newspapers, magazines and books) and additional word norms. By providing a window on quantitative historical psychology, this approach could inform policy and economic history.

Procedure

Question: What influences national wellbeing? Can we produce a historical measure of subjective wellbeing on par with Maddison's historical estimates of GDP?

ENGLISH	VALENCE
aardvark	6.26
abalone	5.3
abandon	2.84
abandonment	2.63
abbey	5.85
abdomen	5.43
abdominal	4.48
abduct	2.42
abduction	2.05
abide	5.52
abiding	5.57
ability	7
abject	4
ablaze	5.15
able	6.64
abnormal	3.53
abnormality	3.05
abode	5.28
abolish	3.84
abominable	4.05
abomination	2.5
abort	3.1
abortion	2.58

ITALIAN	VALENCE
abbaglio	3.94
abbandonato	2
abbondanza	6.82
abbraccio	7.7
abete	6.17
abitante	5.67
abitazione	6.46
abito	7.27
abitudini	4.91
aborto	2.06
abuso	1.74
accettazione	5.79
accogliente	8.03
accomodante	6.4
accordo	6.71
acqua	7.78
adorabile	7.33
adulto	5.78
aereo	6.56
affamato	4.74
affascinare	7.97
affaticato	3.73
affetto	7.48

There are valence norms in many languages

Data: Billions of words of historical natural language (Spanish, French, German, Italian, British English and American English) and affective norms for words in six languages.

Hills, Proto, & Sgroi (2015). Historical analysis of national subjective wellbeing using millions of digitized books. IZA No. 9195

Validating our measure with Eurobarometer data

Unit of analysis: years and languages

Data: Billions of words of historical natural language (German, Italian, British English and American English) and affective norms for words in six languages.

$$NVI = \sum_i v_i p_i$$

mean valence across words

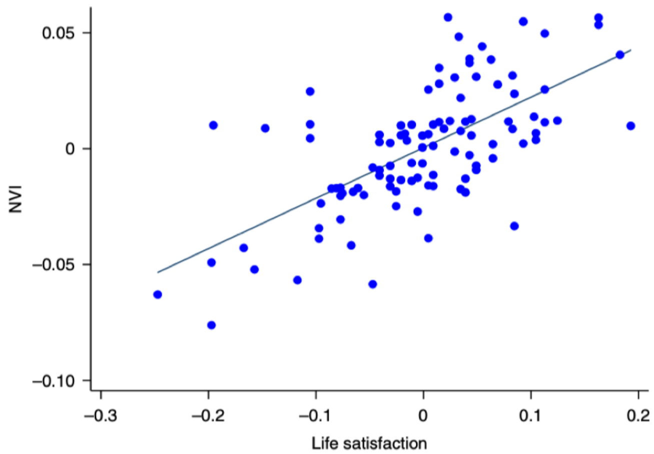


Fig. 1 | Correlation of the NVI and aggregate life satisfaction data from the Eurobarometer survey. The NVI (our measure of subjective wellbeing derived from digitized text) is compared with aggregate life satisfaction (obtained from the Eurobarometer survey-based measure) for the United Kingdom, Germany and Italy (the three countries for which both measures exist) from 1973 to 2009 (the period over which both measures are available). Both variables (the NVI and Eurobarometer life satisfaction measures) are expressed in the form of residuals after controlling for country fixed effects so that values represent variations around the averages for each of the three countries.

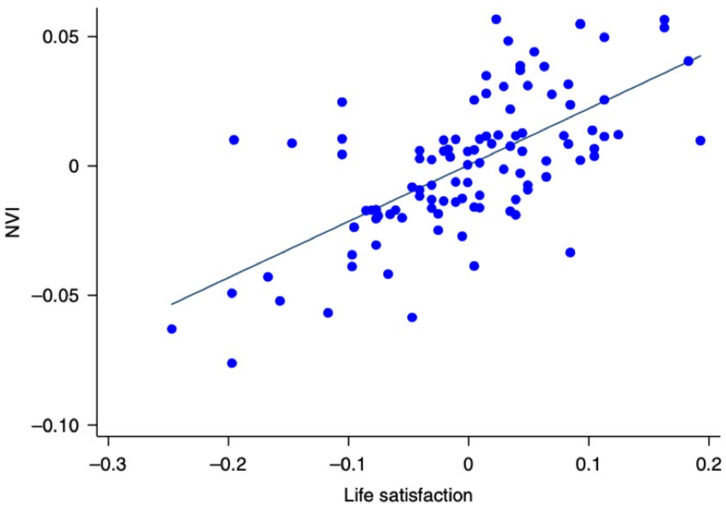


Fig. 1 | Correlation of the NVI and aggregate life satisfaction data from the Eurobarometer survey. The NVI (our measure of subjective wellbeing derived from digitized text) is compared with aggregate life satisfaction (obtained from the Eurobarometer survey-based measure) for the United Kingdom, Germany and Italy (the three countries for which both measures exist) from 1973 to 2009 (the period over which both measures are available). Both variables (the NVI and Eurobarometer life satisfaction measures) are expressed in the form of residuals after controlling for country fixed effects so that values represent variations around the averages for each of the three countries.

Table 1 | The NVI predicts aggregate life satisfaction

	Year fixed effects	**Country-specific trends**
NVI (β (s.e.))	2.8551*** (0.2867)	1.6596** (0.2246)
GDP	Yes	Yes
Country-specific trend	No	Yes
Year fixed effects	Yes	No
r^2	0.730	0.588
n	104	104

The NVI is a statistically significant predictor in an ordinary least squares estimate with country fixed effects of aggregate life satisfaction. The dependent variable is average life satisfaction per country and year, obtained from the Eurobarometer survey-based measure. The period covered is 1973 to 2009, the period over which both measures exist. The countries considered are Germany, Italy and the United Kingdom, the three countries for which both datasets exist. GDP per capita (expressed in terms of purchasing power parity) was obtained from the PWT 8.0 dataset. Column 1 includes year fixed effects (to help to deal with spurious correlations over time) and column 2 includes country-specific trends (to help to deal with spurious correlations across countries). Robust standard errors clustered at country levels are given in brackets. **$P < 0.05$, ***$P < 0.01$. Full statistical information for this table is provided in the Supplementary Information.

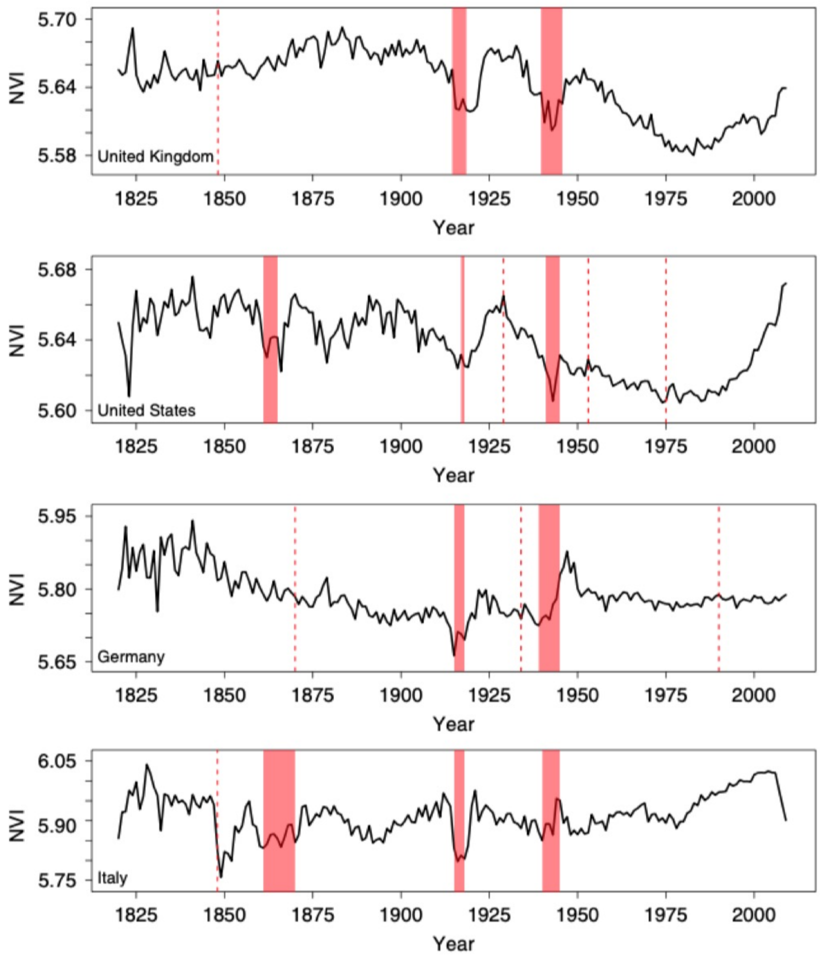


Fig. 2 | NVI through the period 1820–2009. The NVI from 1820 to 2009. Various important events are highlighted in red (for periods of time) or with a dashed vertical red line for events that correspond to a single year. For all countries, the red shaded lines include World War I (approximately 1914–1918) and World War II (approximately 1938–1945). In the three European countries, the line in 1848 indicates the Year of Revolution. In the United States, there is an additional shaded area that represents the Civil War (1861–1865) and vertical red lines that represent the Wall Street Crash (1929), the end of the Korean War (1953) and the fall of Saigon (1975). For Germany, the vertical red lines represent the end of Franco-Prussian War and reunification (1870), Hitler's ascendency to power (1934) and the reunification (1990). In Italy, there is an additional shaded area that represents the unification (1861–1870).

This has a lot of problems, you can probably think of some easily. Let's talk about them.

Word features: A basic tool to get you started

- **The Macrocope**
  - http://macroscope.intelligence-media.com/
- **Textsite**
  - https://warwick.ac.uk/fac/sci/psych/people/thills/thills/textsight/
  - (or search for Hills and TextSight)

3.

Word and document similarities

What are the meanings of the words?

Semantics

- Semantics focuses on word ‘meanings’
- “You shall know a word by the company it keeps.” (Firth 1957)

- In other words, ***semantics is structure.***
- This structure is often referred to now as *embeddings* or *vectors*.
- If we have structure, we can measure similarity.

Question: Do Conspiracy Theorists have a *conspiracy worldview*, apparent in the structure of their language about conspiracies?

LOCO: The 88-million-word language of conspiracy corpus

Alessandro Miani[1] · **Thomas Hills**[2,3] · **Adrian Bangerter**[1]

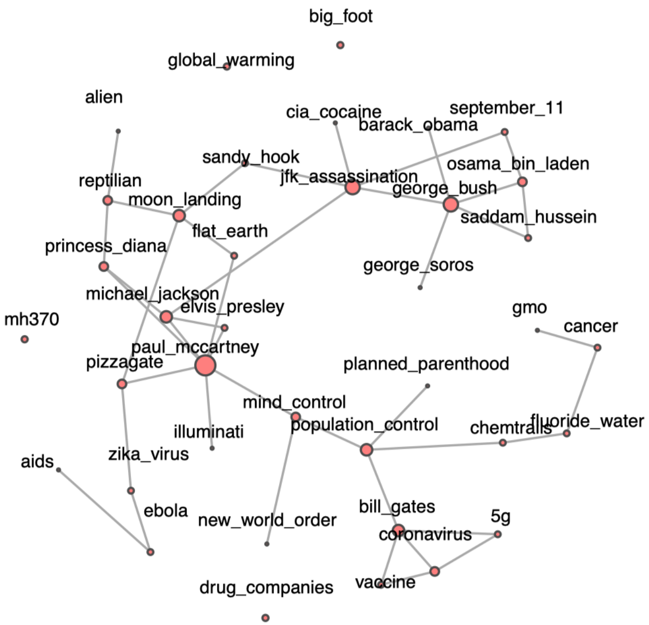
Accepted: 26 August 2021

Abstract

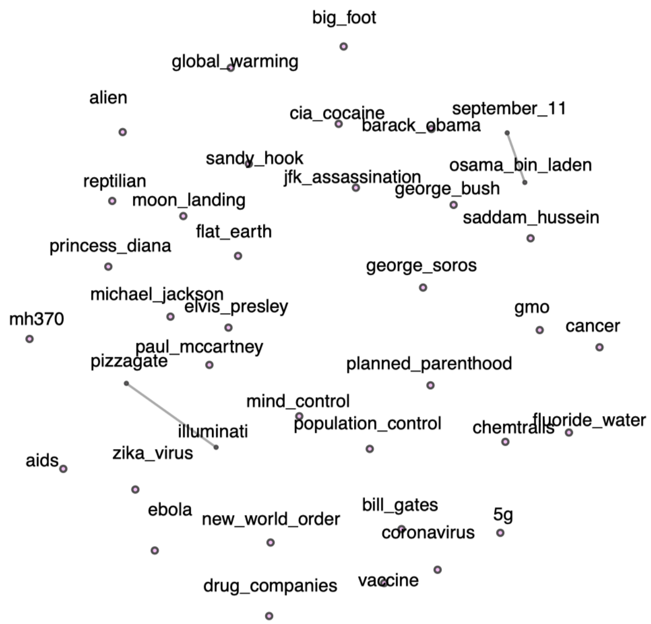
The spread of online conspiracy theories represents a serious threat to society. To understand the content of conspiracies, here we present the language of conspiracy (LOCO) corpus. LOCO is an 88-million-token corpus composed of topic-matched conspiracy (N = 23,937) and mainstream (N = 72,806) documents harvested from 150 websites. Mimicking internet user behavior, documents were identified using Google by crossing a set of seed phrases with a set of websites. LOCO is hierarchically structured, meaning that each document is cross-nested within websites (N = 150) and topics (N = 600, on three different resolutions). A rich set of linguistic features (N = 287) and metadata includes upload date, measures of social media engagement, measures of website popularity, size, and traffic, as well as political bias and factual reporting annotations. We explored LOCO's features from different perspectives showing that documents track important societal events through time (e.g., Princess Diana's death, Sandy Hook school shooting, coronavirus outbreaks), while patterns of lexical features (e.g., deception, power, dominance) overlap with those extracted from online social media communities dedicated to conspiracy theories. By computing within-subcorpus cosine similarity, we derived a subset of the most representative conspiracy documents (N = 4,227), which, compared to other conspiracy documents, display prototypical and exaggerated conspiratorial language and are more frequently shared on Facebook. We also show that conspiracy website users navigate to websites via more direct means than mainstream users, suggesting confirmation bias. LOCO and related datasets are freely available at https://osf.io/snpcg/.

A comparison of worldviews between conspiracy and mainstream documents

Conspiracy



Mainstream



These are connections that occur more often than we expect at random

Dictionary terms

Table 17.1: List of conspiracy topics.

seed
michael jackson
5g
barack obama
saddam hussein
cancer
global warming
coronavirus
moon landing
reptilian
osama bin laden
september 11
cia cocaine
gmo
fluoride water
drug companies
mind control
vaccine
aids
population control
zika virus
new world order
planned parenthood
illuminati
sandy hook
chemtrails
george soros
alien
princess diana
big foot
ebola
flat earth
mh370
bill gates
george bush
jfk assassination
paul mccartney
elvis presley
pizzagate
jonestown suicide

$$PMI = log\frac{P(x,y)}{P(x)P(y)}$$

Question: What happens to our mental lexicon as we age?

Free associations across the lifespan

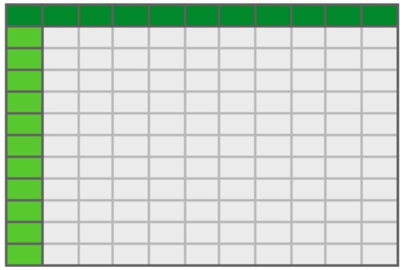
- Study of more than 8000 individuals reporting free associations for 420 words across the lifespan.
- Data is separated into roughly 10-year age groups

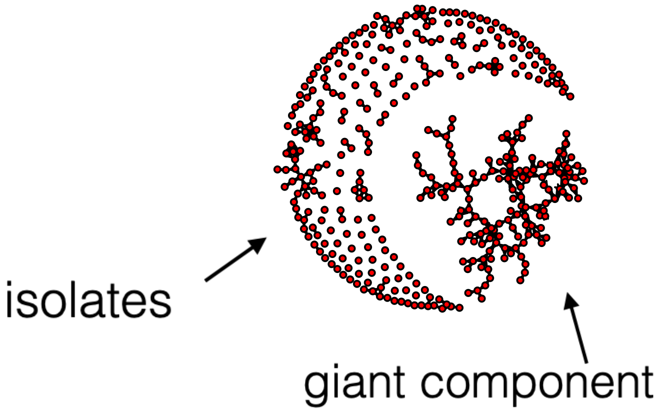
- Study of more than 8000 individuals reporting free associations for 420 words across the lifespan.
- We show people cues, they provide targets.

	Target 1: animal	Target 2: dog	…	Target 3:read
Cue 1:cat	104	53	…	0
Cue 2:book	0	0	…	492
…	…	…	…	…
Cue 420: happy	0	2	…	0

$$w_{ij} = \sum_{i=1}^{P} \frac{w_{i,p}}{N_p - 1}$$

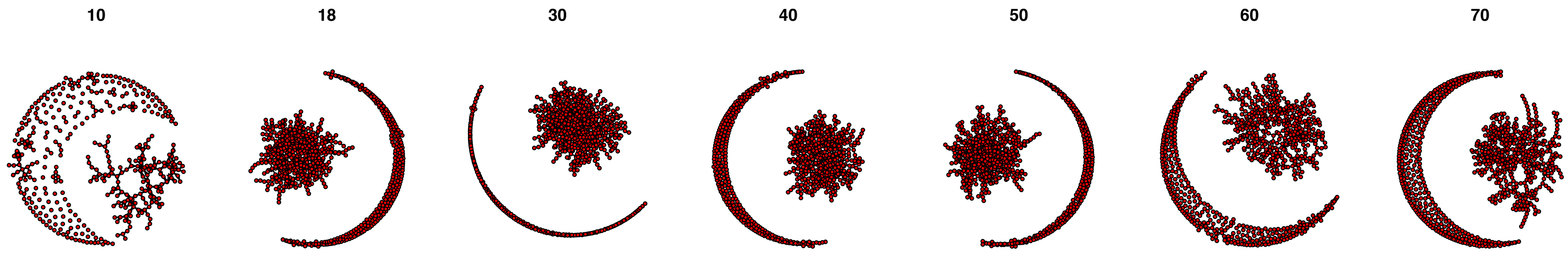
cue x cue matrix





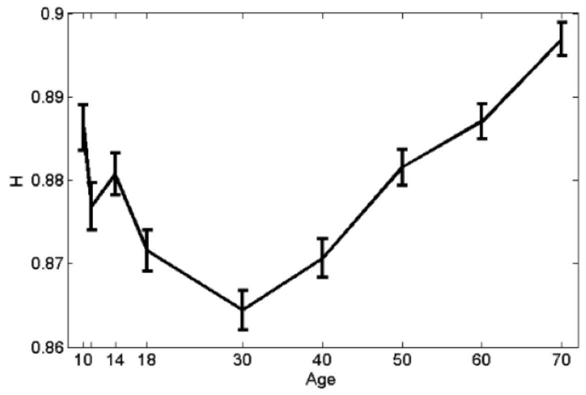
Dubbosarsky, De Deyne, and Hills (2017). *Developmental Psychology*

Free associations across the lifespan



Networks of free associations

$$H = -\sum_{i=1}^{n} \frac{p(x_i)\log_b(p(x_i))}{\log_b(n)}$$



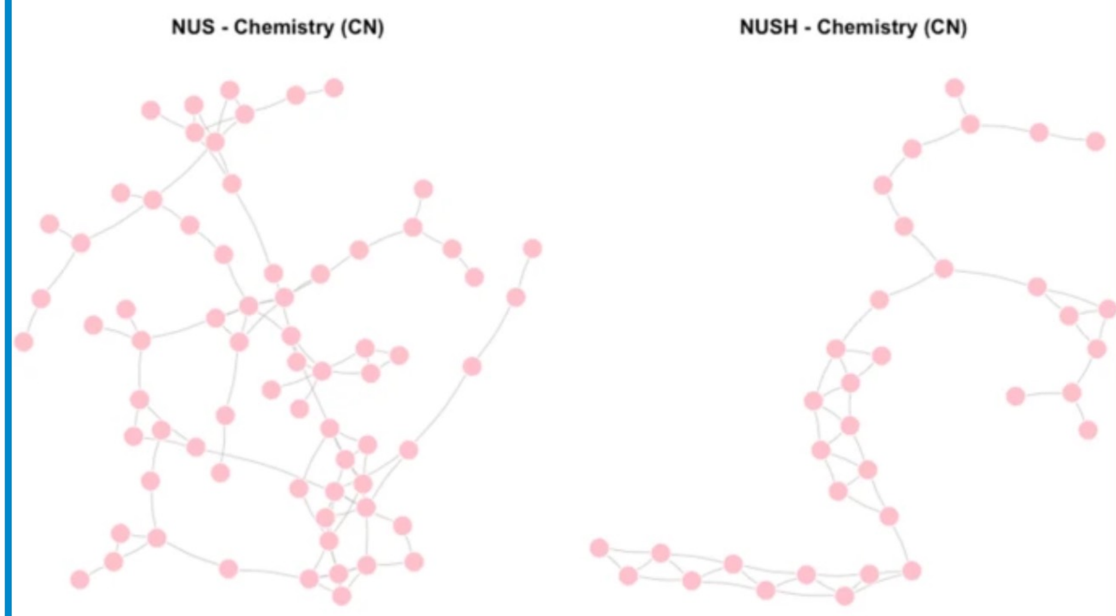
Dubossarsky, Hills, & De Deyne, 2017

Two approaches to semantics

- Free associations—what do people say it means.
- Semantic space models, or vector-based semantics (BEAGLE, HAL, LSA, Word2Vec, BERT, RoBERTa, GPT#)—What are meaning relations embedded in natural language?

Fluency task comparisons (novice vs. experts)

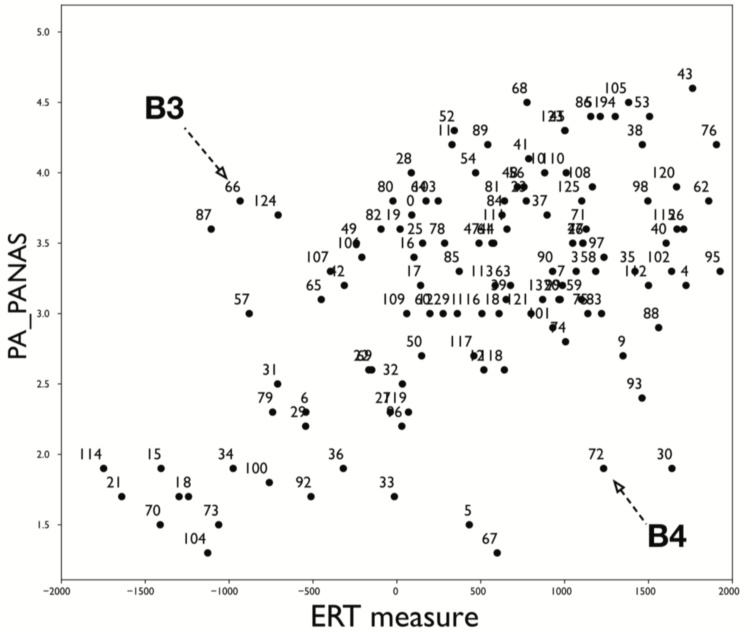
- Siew & Guru (2023).
- Comparison of National University of Singapore undergrads versus NUS high school students.
- ‘Fluency’ tasks invite participants to say what comes to mind when they think about ‘X’.



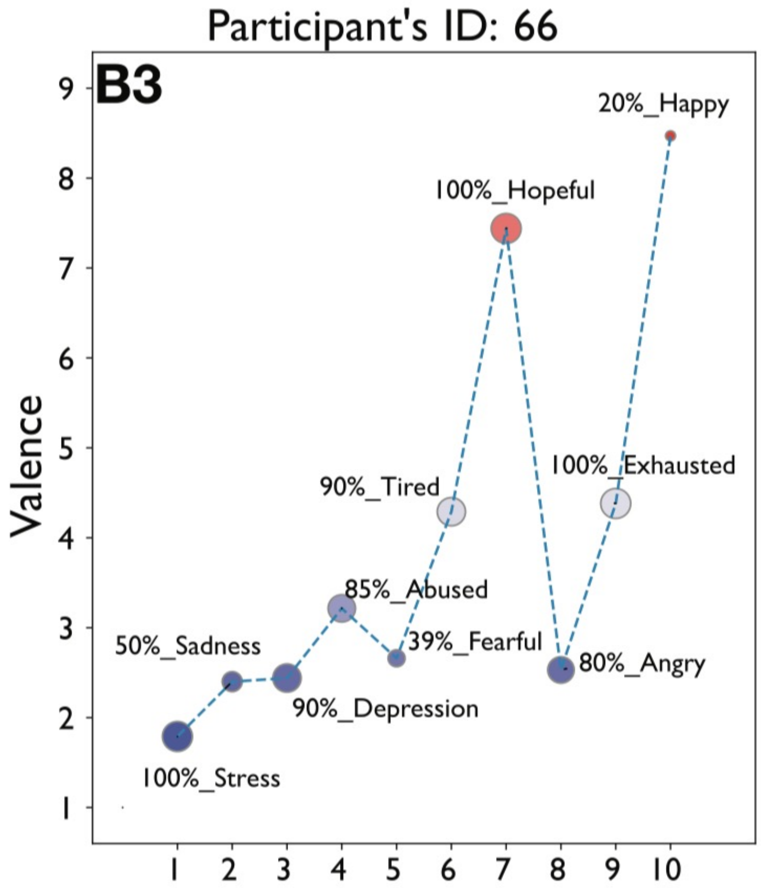
Emotional Recall Task (Li et al., 2020)

- "Tell me 10 emotions you've felt recently"

Survey measure



Our measure



Two approaches to semantics

- Free associations—what do people say it means.
- Semantic space models, or vector-based semantics (BEAGLE, HAL, LSA, Word2Vec, BERT, RoBERTa, GPT#)—What are meaning relations embedded in natural language?

Semantic space models: Example

- Term-document matrix
- Usually involves inverse-document frequency to account for word specificity

$$g_i = \log_2 \frac{n}{1 + \mathrm{df}_i}$$

```
"The man walked the dog"
"The man took the dog to the park"
"The dog went to the park"

You prepare a matrix of word counts like so:
```

	Passage 1	Passage 2	Passage 3
the	2	3	2
man	1	1	0
walked	1	0	0
dog	1	1	1
took	0	1	0
to	0	1	1
park	0	1	1
went	0	0	1

Latent Semantic Analysis (LSA)

LSA is essentially low-rank *approximation* of document term-matrix

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

term-document matrix
6 terms, 4 documents

=

Word assignment to topics

	IT	cars
linux	-0.33	-0.53
modem	-0.32	-0.54
the	-0.62	-0.10
clutch	-0.38	0.42
steering	-0.36	0.25
petrol	-0.37	0.42

X

Topic Importance

11.4	
	6.27

We can change this dimension

X

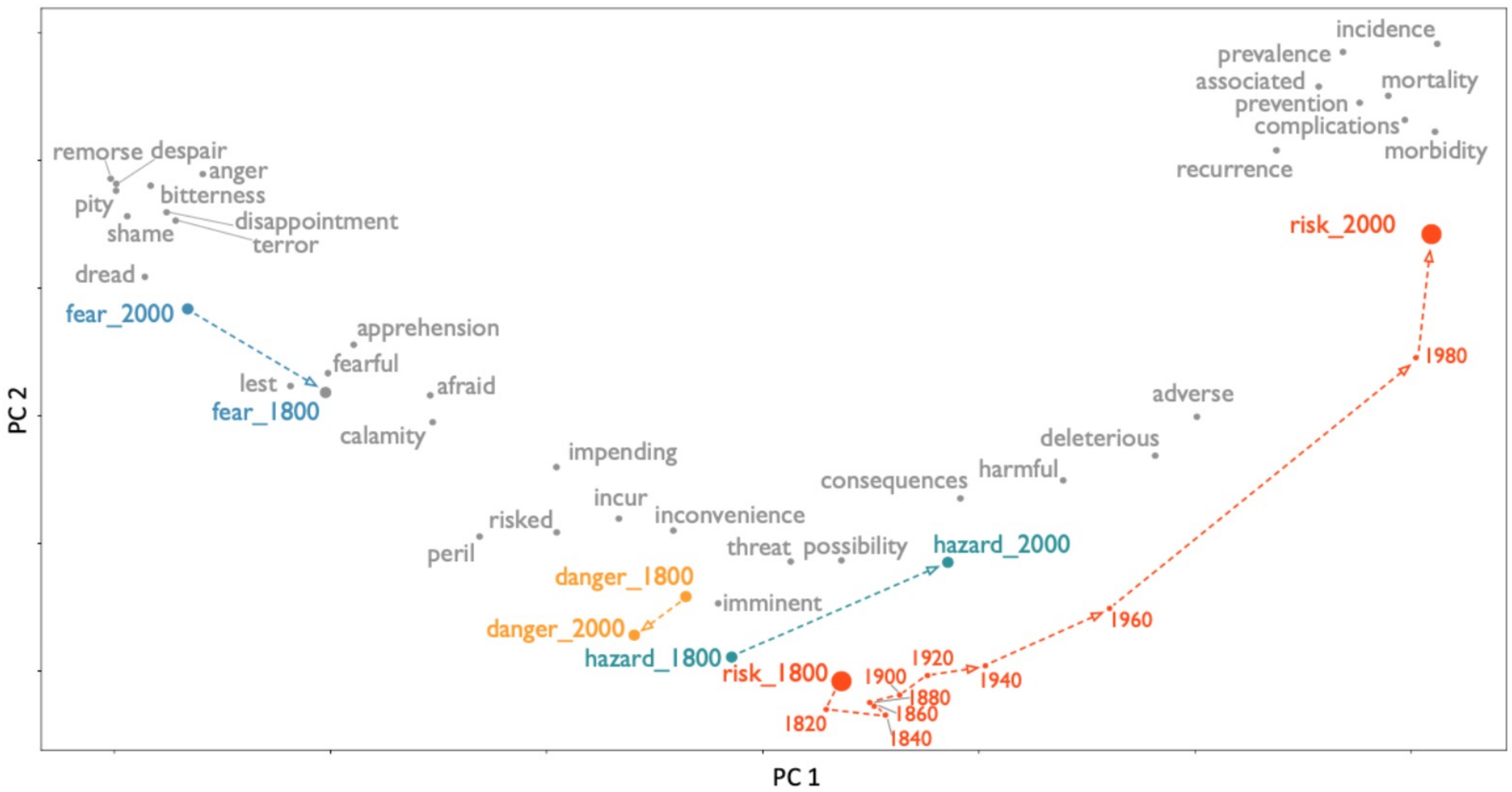
Topic distribution acr documents

	D1	D2	D3	D4
IT	-0.42	-0.48	-0.57	-0.51
cars	-0.56	-0.52	0.45	0.46

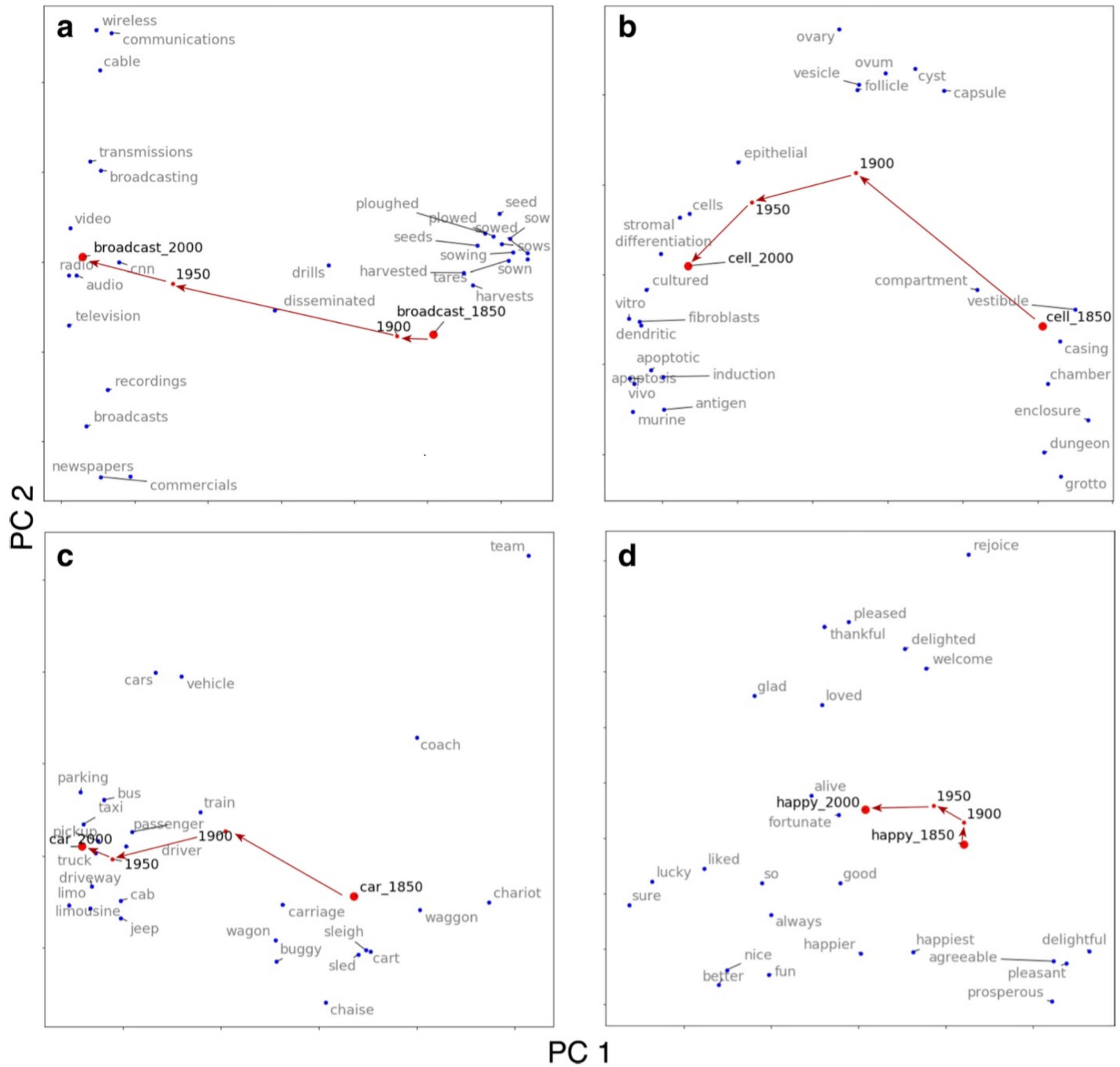
Singular-value decomposition: A matrix algebra method

Did risk change its meaning over the last 200 years?

PC = principal components analysis – dimensionality reduction method



Changes in word meanings (Li et al., 2019)



PC = principal components analysis dimensionality reduction method

4.

Topic modelling

What topics do the words refer to?

Question: What are people talking about when they're talking about 'immigrants'?

Cognition 215 (2021) 104813

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Language patterns of outgroup prejudice

Ying Li [a,*], Thomas T. Hills [b]

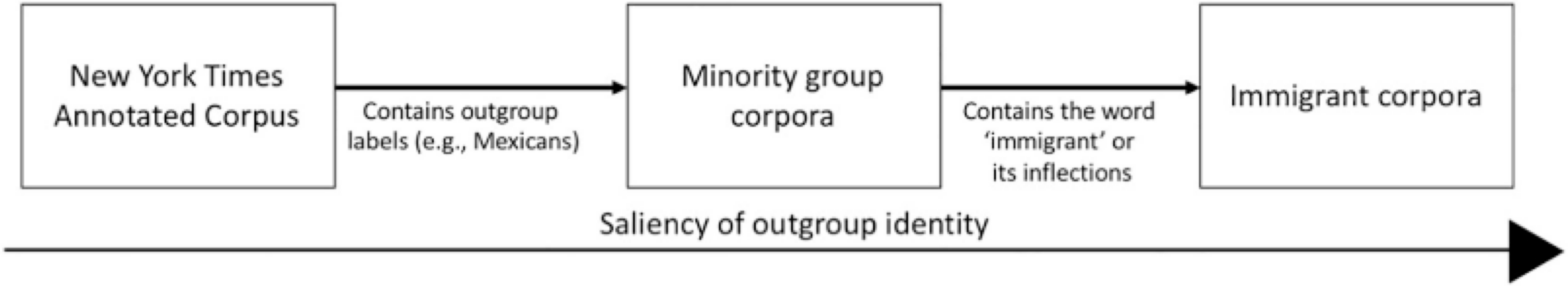
[a] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[b] Department of Psychology, University of Warwick, UK

The approach

1. Is language more concrete around social groups that have closer perceived social distance? (construal level theory)
2. Are minority groups with more abstract language described with more negative language? (linguistic expectancy theory)
3. What are the topics associated with explicit references to immigration?
4. How are the topics distributed across minority groups?
5. How are immigrant topics associated with sentiment?

The approach

- To evaluate these questions we used the New York Times Corpus available from the Linguistic Data Consortium.
- 1.8 million articles, published from 1981 to 2007.
- From this we built an immigrant corpus for ~60 immigrant groups



Cognition 215 (2021) 104813

Contents lists available at ScienceDirect

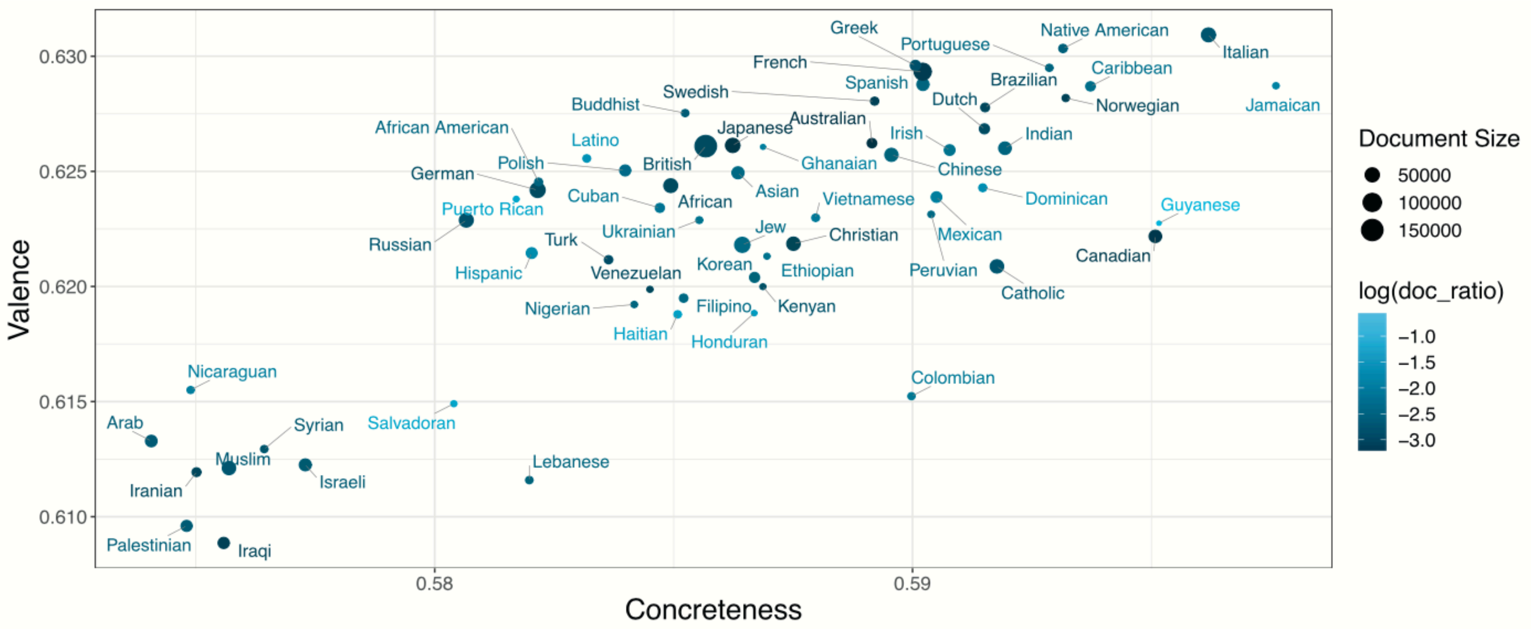
Cognition

journal homepage: www.elsevier.com/locate/cognit

Language patterns of outgroup prejudice

Ying Li [a,*], Thomas T. Hills [b]

[a] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[b] Department of Psychology, University of Warwick, UK



Topic modelling or Latent Dirichlet Allocation

- Unsupervised categorization algorithm—a clustering algorithm.
- LDA assumes that documents are *made up of multiple topics* and *topics are made up of multiple words*.
- To create a document, the model chooses a distribution over topics (70% Psychology, 20% Linguistics, 10% Star Wars)
- To identify topics from a document, we need to reverse this process to find the hidden structure.

Perspectives on Psychological Science
1–8

DOI: 10.1177/1745691618803647
www.psychologicalscience.org/PPS
SAGE

The Dark Side of Information Proliferation

Thomas T. Hills
Department of Psychology, University of Warwick

Abstract

There are well-understood psychological limits on our capacity to process information. As information proliferation—the consumption and sharing of information—increases through social media and other communications technology, these limits create an attentional bottleneck, favoring information that is more likely to be searched for, attended to, comprehended, encoded, and later reproduced. In information-rich environments, this bottleneck influences the evolution of information via four forces of cognitive selection, selecting for information that is belief-consistent, negative, social, and predictive. Selection for belief-consistent information leads balanced information to support increasingly polarized views. Selection for negative information amplifies information about downside risks and crowds out potential benefits. Selection for social information drives herding, impairs objective assessments, and reduces exploration for solutions to hard problems. Selection for predictive patterns drives overfitting, the replication crisis, and risk seeking. This article summarizes the negative implications of these forces of cognitive selection and presents eight warnings that represent severe pitfalls for the naive "informavore," accelerating extremism, hysteria, herding, and the proliferation of misinformation.

To produce a document the author chooses topics and then they choose words from those topics.

The output of the model is a distribution of topics per document

And

A distribution of words per topic.

So...What are people talking about when they're talking about 'immigrants'?

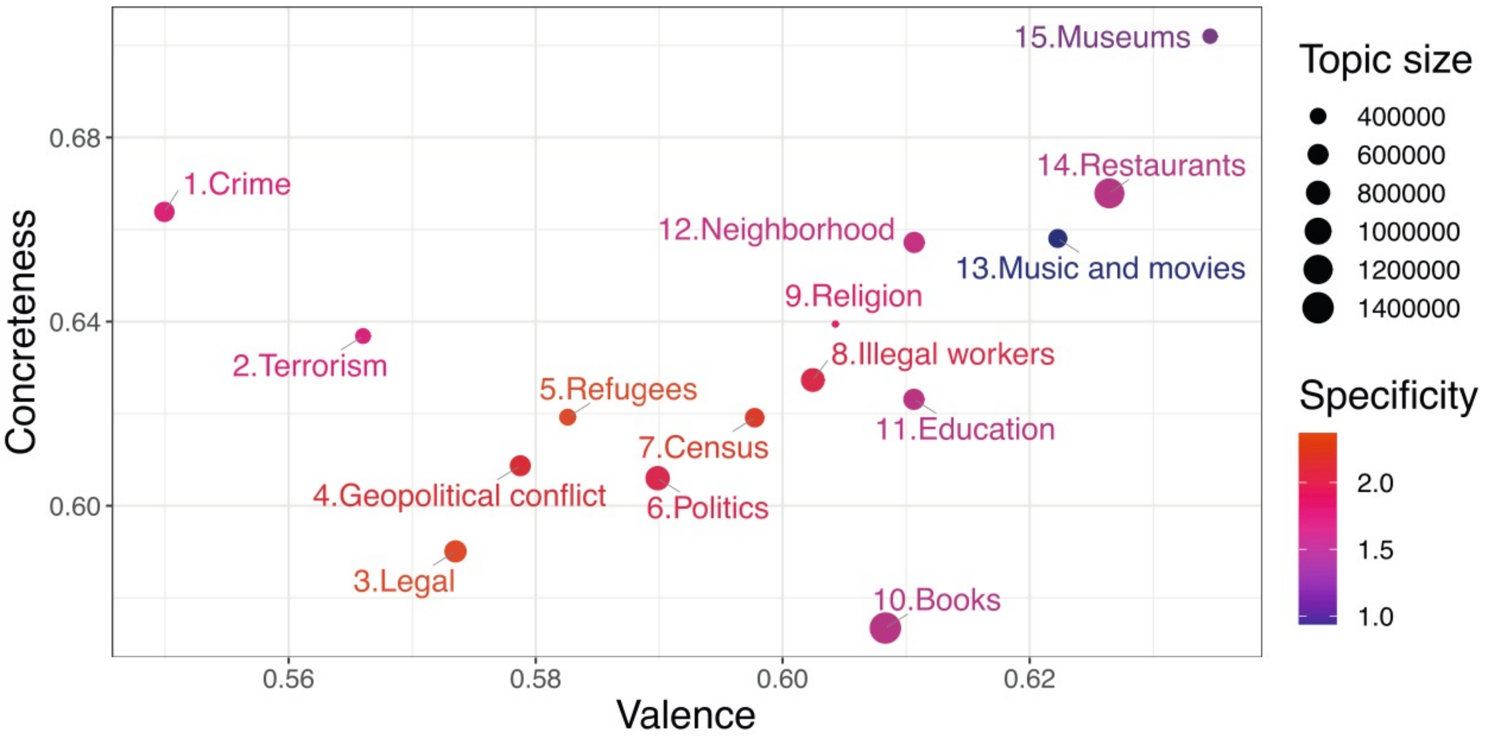
Topics in the immigrant corpora

Table 3
Top 10 keywords for each topic (from most negative to most positive).

Index	Topic	Keywords
1	Crime	police, officer, arrest, charge, prosecutor, drug, kill, gang, crime, shoot.
2	Terrorism	Muslim, terrorist, bomb, attack, intelligence, Islamic, FBI, mosque, Sept, Iraq
3	Legal	immigration, law, court, alien, judge, legal, justice, case, federal, lawyer
4	Politics	Republican, Bush, Democrat, bill, president, vote, senate, senator, campaign, Clinton
5	Geopolitical conflict	Israel, minister, Soviet, France, Germany, Europe, party, prime, Palestinian, Jew
6	Refugees	refugee, Cuban, asylum, Haitian, unite, Miami, boat, Castro, state, official
7	Illegal workers	worker, border, Mexican, company, labor, job, wage, work, pay, illegal
8	Census	Hispanic, population, percent, Asian, Black, census, Chinese, Korean, Latino, immigrant
9	Neighborhood	city, build, house, neighborhood, county, resident, island, apartment, rent, community
10	Books	write, book, life, American, world, think, history, story, time, way
11	Religion	church, Catholic, Irish, bishop, priest, Jewish, religious, parish, pope
12	Education	school, student, child, teacher, education, parent, program, health, care, college
13	Restaurants	restaurant, cook, eat, chicken, room, shop, soccer, dish, food, cup
14	Music & movies	theater, film, music, movie, play, art, direct, musical, dance, song, artist
15	Museums	museum, Sunday, tour, street, information, tomorrow, admission, exhibition, park, sponsor

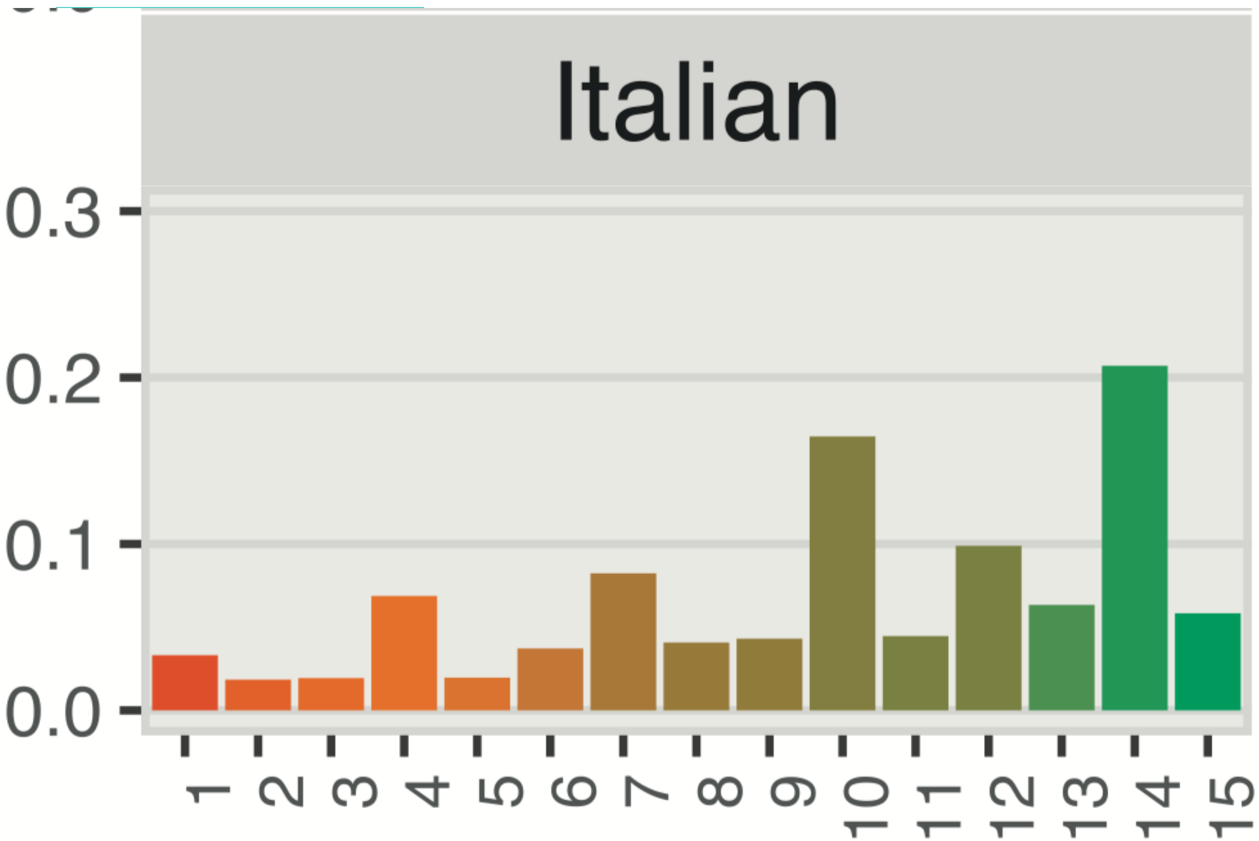
We combined inflections (e.g., *German*, *Germany*) to avoid unnecessary duplications. An interactive visualization of topic–word association with varying degrees of lambda can be accessed at https://liyingpsych.github.io/LanguageOfPrejudice/. The visualization was generated by R package LDAvis (Sievert & Shirley, 2014). Lambda was set to 0.3 when displaying keywords for topic 13 (Restaurants) because this topic was mixed with generic linguistic patterns underlying all articles (e.g., *say, like, one, day, get, come*). Reducing lambda further penalizes the weight of high frequency words that tend to appear across all articles.

Apply feature analysis to topics



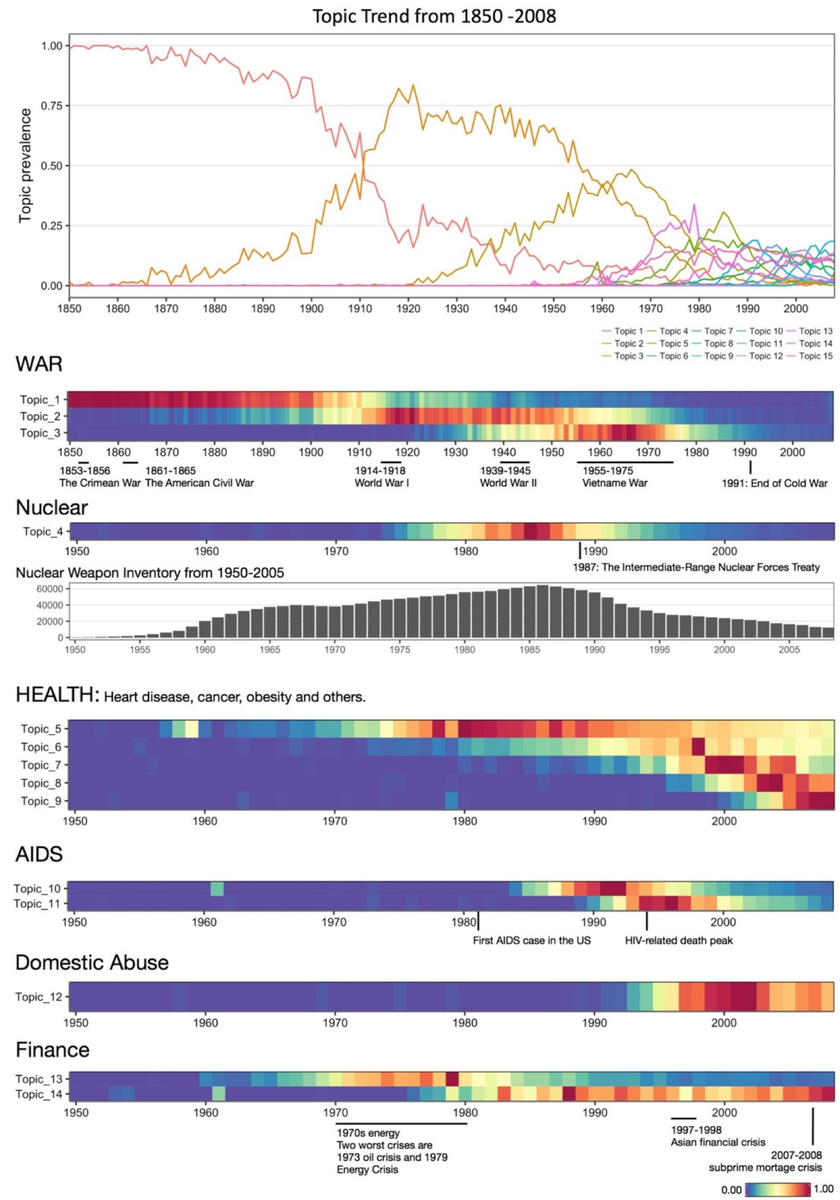
Specificity: a measure of how likely people are to be talking about immigrants when this topic is present.

What groups are associated with which topics?



Question: What were people talking about when they were talking about 'risk' in the past?

Topic modelling of risk



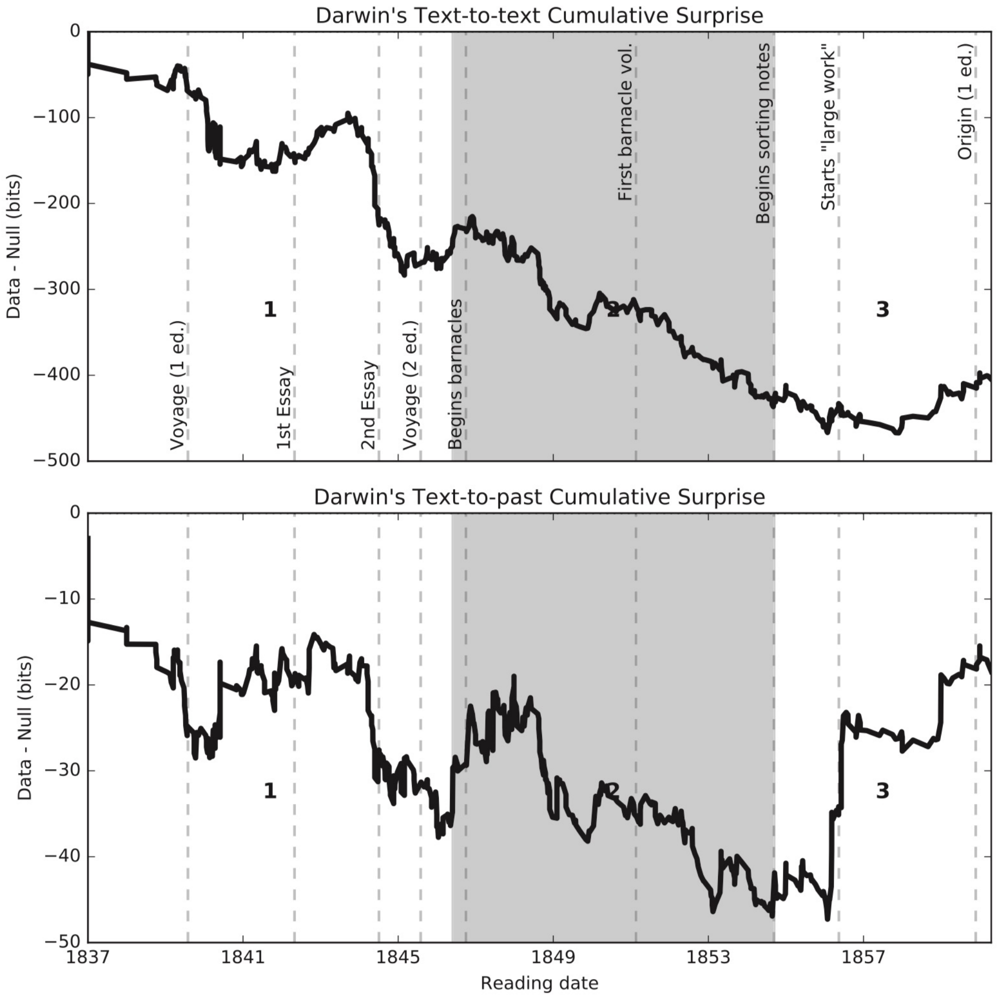
Question: When did Darwin explore and exploit in his reading?

Darwin’s reading

Murdock, Allen, DeDeo 2017

Table 1
Timeline. Major events in Charles Darwin’s life, including those marked in Fig. 1. This paper focuses on the critical period of his work from 1837 to 1860, leading to the publication of *The Origin of Species* (boundaries marked in bold). See Berra (2009) for an expanded chronology.

	Major Events in Charles Darwin’s Life (1809–1882)
12 February 1809	Born in Shrewsbury, England
22 October 1825	Matriculates at University of Edinburgh
15 October 1827	Admitted to Christ’s College, Cambridge
27 December 1831	Departs England aboard the *HMS Beagle*
2 October 1836	Return to England aboard the *HMS Beagle*
July 1837	**First entries in reading notebooks**
August 1839	Publication of *The Voyage of the Beagle* (1st edition)
May 1842	Writes the 1st Essay on Species
4 July 1844	Writes the 2nd Essay on Species
August 1845	Publication of *The Voyage of the Beagle* (2nd edition)
1 October 1846	Begins barnacle project
19 February 1851	Publishes first volume of barnacle work
9 September 1854	Begins sorting notes on natural selection
14 May 1856	Starts writing “large work” on species
24 November 1859	Publication of *The Origin of Species* (1st edition)
13 May 1860	**Last entry in reading notebooks**
24 February 1871	Publication of *The Descent of Man*
19 February 1872	Publication of *The Origin of Species* (6th and final edition)
21 April 1882	Dies at Down House in Kent, England



Main approaches in NLP

1. Counting words
2. Word feature analysis (sentiment)
3. Word and document similarity
4. Topic modelling

5.

General Questions for NLP

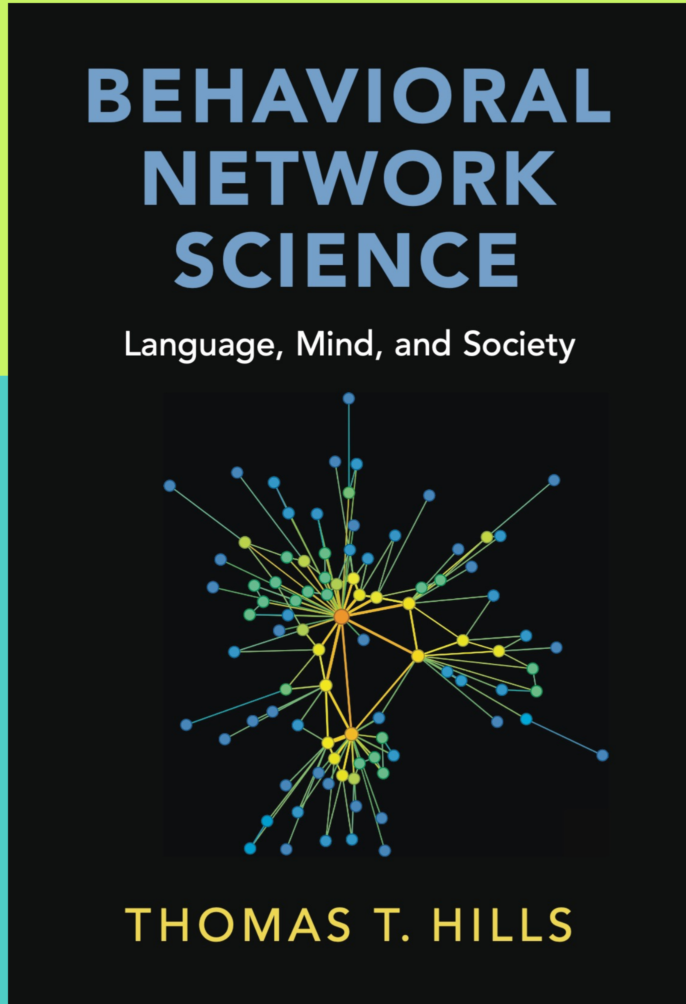
Limitations

Frequent limitations in NLP

- What's the *data quality*?
- Is the data *representative*?
- Are there *Independent corpora*?
- What are the *alternative hypotheses*? Are they tested?
- Do we need statistics for this? Not always.
- How big is the effect? Measure it relative to something you already understand.

Thank you

Questions?



Cambridge University Press