The baby-boomers' boom: trends in semantic changes in American English

Søren Wichmann
Leiden University

[Automatic Detection of Language Change
(workshop adjacent fo SLTC)
Stockholm, Nov. 7, 2018]

Structure of the presentation

- Data: a sample of Corpus of Historical American English (COHA)
- Methods: different vector semantic approaches
- A methodological result: sensitivity of different metrics to frequencies
- An empirical result from the case study: average semantic change in the core lexicon driven by words relating to individualism
- Explaining the observations with reference to social psychology

Corpus of Historical American English (COHA)

http://corpus.byu.edu/coha/

- 20 decades: from 1810-1819 to 2000-2009
- ~400 M words
- Balanced with respect to genres
- Corpus used here:
  - 2-grams , case insensitive, with POS tags
- In order to focus on basic vocabulary only the 23,519 word forms that appear in all decades in the 2-gram corpus are used (mysteriously not exactly the same set as words that appear in all decades in the 1-gram corpus)

Methods

- I use different vector semantic measures and initially remain agnostic about the appropriateness of each:
  - Jensen-Shannon
  - Jaccard
  - Manhattan
  - cosine
  - correlation
  - Euclid
  - Chebyshev
  - simple distribution measure (SDM)
- Using each metric, the semantic distance between a word form and itself between adjacent decades is computed

Jensen-Shannon divergence

- Represents the divergence of each distribution from the mean of the two

$$JS(q,r) = \frac{1}{2}\left[D\left(q \,\middle\|\, \text{avg}_{q,r}\right) + D\left(r \,\middle\|\, \text{avg}_{q,r}\right)\right] \text{ , where}$$

$$D(p_1(V) \,\|\, p_2(V)) = \sum_v p_1(v) \log \frac{p_1(v)}{p_2(v)}$$

(Lee 1999: 26)

Jaccard's coefficient

- Computes a weighted number of overlapping features

$$\mathrm{Jac}(q,r) = \frac{|\{v : q(v) > 0 \text{ and } r(v) > 0\}|}{|\{v \mid q(v) > 0 \text{ or } r(v) > 0\}|}$$

(Lee 1999: 27)

Manhattan distance

- The sum of absolute differences between distributions

$$L_1(q,r) = \sum_v |q(v) - r(v)|$$

(Lee 1999: 27)

Cosine

$$\cos(q,r) = \frac{\sum_v q(v)r(v)}{\sqrt{\sum_v q(v)^2}\sqrt{\sum_v r(v)^2}}$$

(Lee 1999: 27)

Example:

	large	data	computer
apricot	2	0	0
digital	0	1	2
information	1	6	1

(Jurafsky and Martin 2016ms)

$$\cos(\text{apricot},\text{information}) = \frac{2+0+0}{\sqrt{4+0+0}\sqrt{1+36+1}} = \frac{2}{2\sqrt{38}} = .16$$

Correlation

- Simply a Pearson correlation of two (normalized) frequency vectors

Euclid

$$L_2(q,r) = \sqrt{\sum_v (q(v) - r(v))^2}$$

(Lee 1999: 27)

Chebyshev

The maximal distance between the values of two frequency vectors

$$D_{\text{Chebyshev}}(p, q) := \max_i(|p_i - q_i|)$$

(https://en.wikipedia.org/wiki/Chebyshev_distance)

Example:

The Chebyshev Distance between point A and B is

$$d_{BA} = \max\{|0-7|, |3-6|, |4-3|, |5+1|\}$$
$$= \max\{7, 3, 1, 6\} = 7$$

(from Teknomo's website)

Simple distribution measure (SDM)

- Invented by Vladimir Bochkarev, Kazan Federal University
- Measures the sum of relative frequencies of co-occurrences in bigrams between A and words not co-occurring with B plus the sum of relative frequencies of co-occurrences in bigrams between B and words not co-occurring with A
- When divided by four this produces a number between 0 and 1, where the former means that all contexts are shared and the latter that no contexts are shared

invented example:
word A: ‘great’, word B: ‘last’

xA	freq(xA)	Ax	freq(xA)	xB	freq(xB)	Bx	freq(Bx)
her	14	smile	20	her	12	affair	5
a	18	laugh	3	a	16	state	4
little	5	request	5	little	6	idea	3
sweet	3	wish	14	cute	5	thing	23
very	6	look	2	great	7	look	5
too	2	desire	8	funny	3	desire	11
your	1	deal	1	one	2	deal	2

invented example:
word A: ‘great’, word B: ‘last’

xA	freq(xA)	Ax	freq(xA)	xB	freq(xB)	Bx	freq(Bx)
her	0.2857	smile	0.3774	her	0.2353	affair	0.0943
a	0.3673	laugh	0.0566	a	0.3137	state	0.0755
little	0.1020	request	0.0940	little	0.1176	idea	0.0566
sweet	0.0612	wish	0.2642	cute	0.0980	thing	0.4334
very	0.1224	look	0.0377	great	0.1372	look	0.0943
too	0.0408	desire	0.1509	funny	0.0588	desire	0.2075
your	0.0204	deal	0.0189	one	0.0392	deal	0.0377

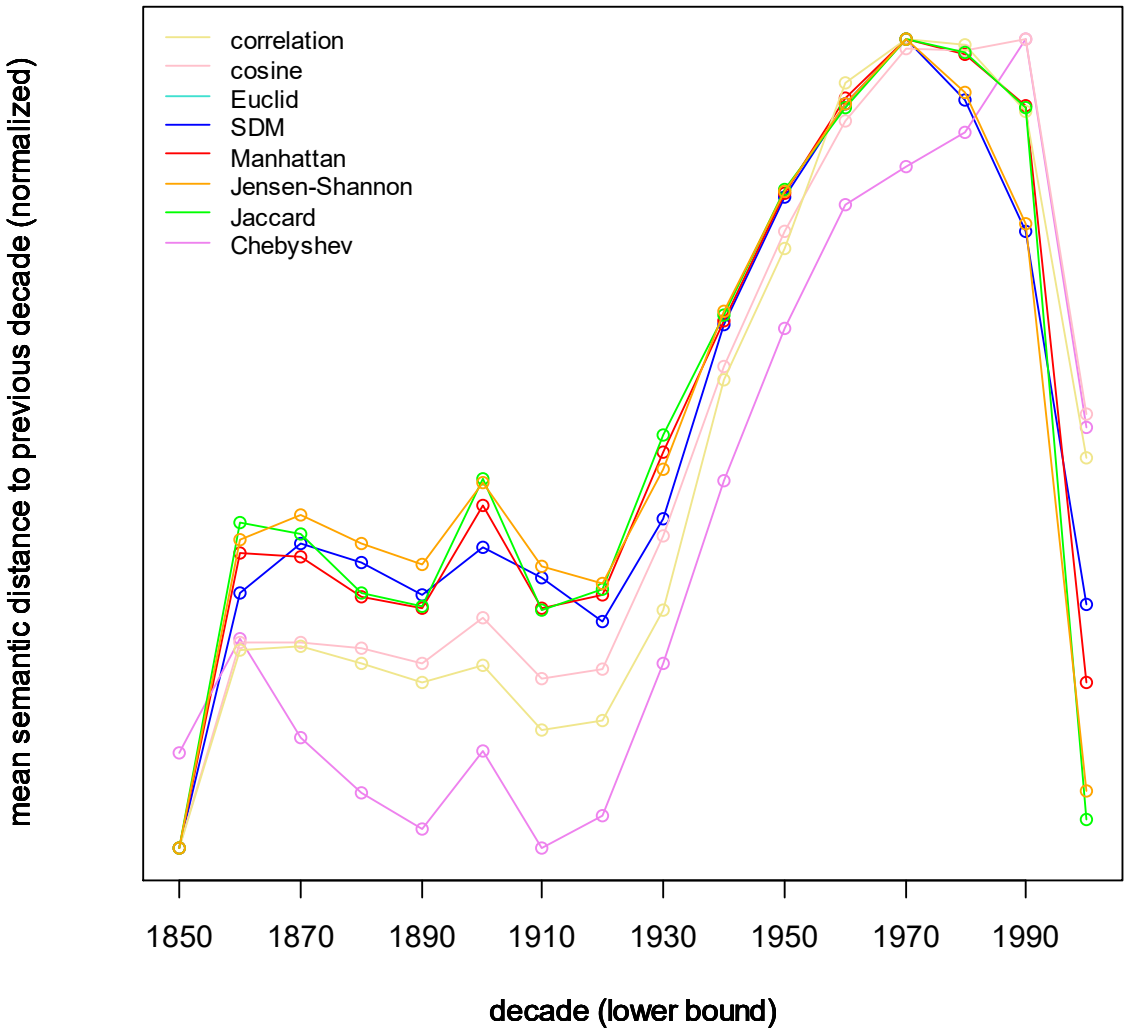
invented example:
word A: ‘great’, word B: ‘last’

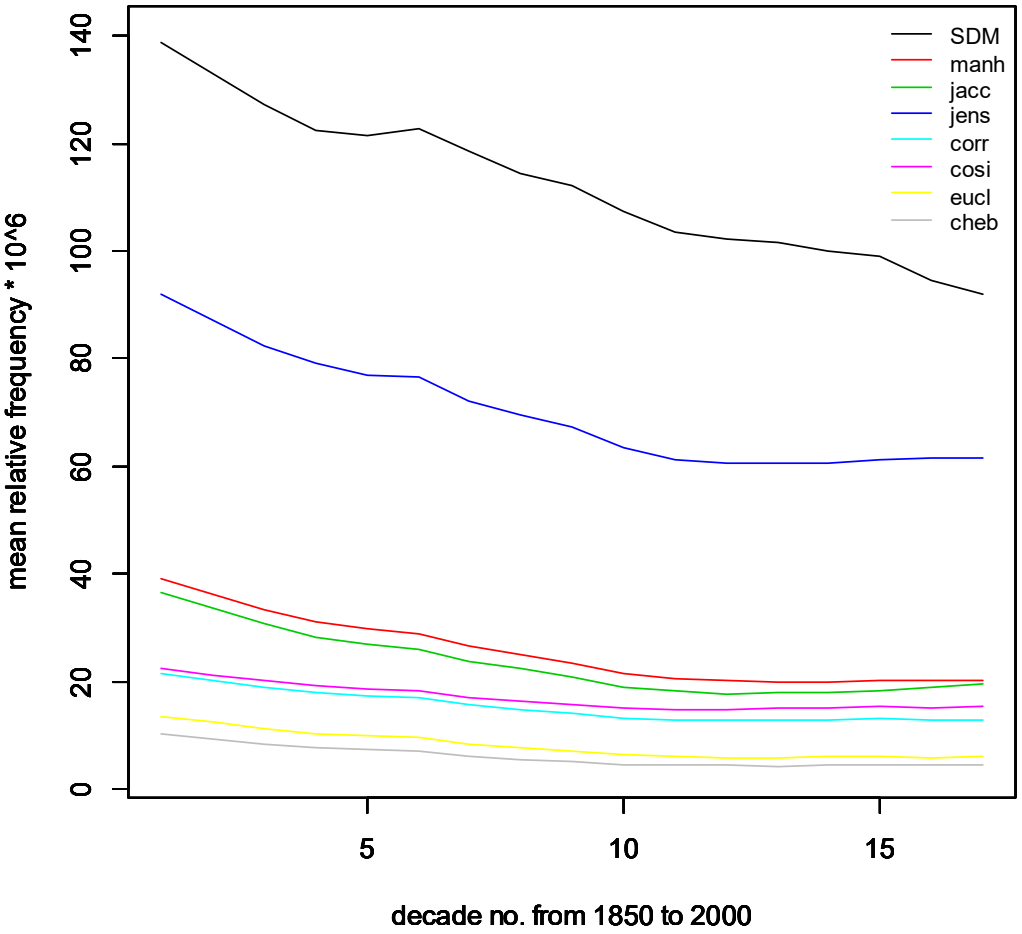
xA	freq(xA)	Ax	freq(xA)	xB	freq(xB)	Bx	freq(Bx)
~~her~~	~~0.2857~~	smile	0.3774	~~her~~	~~0.2353~~	affair	0.0943
~~a~~	~~0.3673~~	laugh	0.0566	~~a~~	~~0.3137~~	state	0.0755
~~little~~	~~0.1020~~	request	0.0940	~~little~~	~~0.1176~~	idea	0.0566
sweet	0.0612	wish	0.2642	cute	0.0980	thing	0.4334
very	0.1224	~~look~~	~~0.0377~~	great	0.1372	~~look~~	~~0.0943~~
too	0.0408	~~desire~~	~~0.1509~~	funny	0.0588	~~desire~~	~~0.2075~~
your	0.0204	~~deal~~	~~0.0189~~	one	0.0392	~~deal~~	~~0.0377~~

invented example:
word A: ‘great’, word B: ‘last’

xA	freq(xA)	Ax	freq(xA)	xB	freq(xB)	Bx	freq(Bx)
~~her~~	~~0.2857~~	smile	0.3774	~~her~~	~~0.2353~~	affair	0.0943
~~a~~	~~0.3673~~	laugh	0.0566	~~a~~	~~0.3137~~	state	0.0755
~~little~~	~~0.1020~~	request	0.0940	~~little~~	~~0.1176~~	idea	0.0566
sweet	0.0612	wish	0.2642	cute	0.0980	thing	0.4334
very	0.1224	~~look~~	~~0.0377~~	great	0.1372	~~look~~	~~0.0943~~
too	0.0408	~~desire~~	~~0.1509~~	funny	0.0588	~~desire~~	~~0.2075~~
your	0.0204	~~deal~~	~~0.0189~~	one	0.0392	~~deal~~	~~0.0377~~
SUM	0.2449		0.7925		0.3333		0.6604

SDM = (0.2449 + 0.7925 + 0.3333 + 0.6604)/4 = 0.5078


correlation
cosine
Euclid
SDM
Manhattan
Jensen-Shannon
Jaccard
Chebyshev
mean semantic distance to previous decade (normalized)
1850
1870
1890
1910
1930
1950
1970
1990
decade (lower bound)



Mean relative frequencies of words whose change curves correlate best ($p < 0.001$) with the average

Sample from 100 words whose change curves correlate best with the mean - Chebyshev

- Endeavors and objectives of the individual, positive qualities
  - vouchsafed, gloried, mirth, amuse, godliness, delightful, kindly, tenderness, merry, benefactions, excellence, congratulations
- Troubles that the individual may encounter, negative qualities
  - ruinous, toiled, laboring, deceive, infest, reproach, abysses, lamentable, mischief, weeping, fretful, criticize, insensible, feeble

Sample from 100 words whose change curves correlate best with the mean - Euclid

- Endeavors and objectives of the individual, positive qualities
  - gloried, delightful, jest, happiness, inspiring, vouchsafed, amuse, priceless, humanity, freest, bravely, endeavor, plentifully, sympathy, entertaining, excellence, wisely
- Troubles that the individual may encounter, negative qualities
  - voraciously, villainous, quell, deceive, toiled, bumpkin, upbraided, roused, laboring, fretful, appeal, toiling, fortune, reproach, compel, savage

Sample from 100 words whose change curves correlate best with the mean - correlation

- Endeavors and objectives of the individual, positive qualities
  - betrothed, entrancing, consummate, invigorated, blessedness, plentifully, splendours, hardihood, jesting, meriting, frugally, jest, amuse, munificent, sweetest, emancipate, dearer, satiated, amusements, sympathizing, gratifies, glutted, effervescence, eventful
- Troubles that the individual may encounter, negative qualities
  - peevishly, pillage, decays, skulking, marauder, afflicted, subverted, bashfulness, endure, insultingly, calamity, toiled, toil, haunts

Sample from 100 words whose change curves correlate best with the mean - cosine

- Endeavors and objectives of the individual, positive qualities
  - emancipate, blessedness, tender-hearted, plentifully, betrothed, frolicsome, amuse, excite, sweetest, happiness, delightful, wish, loftiest, endeavors, exploit, jest, civilized
- Troubles that the individual may encounter, negative qualities
  - voraciously, skulking, flout, simpletons, incommunicable, bashfulness, toiled, mischief, haunts, extricating, endure, quibbles, tediously, laboring, extravagant

Sample from 100 words whose change curves correlate best with the mean - jaccard

- Endeavors and objectives of the individual, positive qualities
  - valorous, splendours, sociably, blessedness, entrancing, trusting, devoutly, humanity, excellence, devotedly, gratifying, invigorated, betrothed, clear-sighted
- Troubles that the individual may encounter, negative qualities
  - voraciously, flouts, skulking, inebriated, simpletons, pander, marauder, laboring displeasing, bashfulness, slander, perplexities, reproaches, labored, desolation, appeal, blighted, necessaries, devilish, raillery

Sample from 100 words whose change curves correlate best with the mean – Manhattan

- Endeavors and objectives of the individual, positive qualities
  - blessedness, plentifully, excellence, humanity, devoutly, devotedly, fancying, jest, fortune, gratifying, emancipate, fancied, righteous, inspiration, praiseworthy, sweetest
- Troubles that the individual may encounter, negative qualities
  - voraciously, predestined, laboring, desolation, worthless, displeasing, reproaches, raillery, wish, solicitations, appeal, perplexities, entrapped, displeases, offence, perishing, reckless, ruin

Sample from 100 words whose change curves correlate best with the mean – Jensen-Shannon

- Endeavors and objectives of the individual, positive qualities
  - frugally, devoutly, gay, convenience, devotedly, helper, praiseworthy, passion, agreeably, jest, fancying
- Troubles that the individual may encounter, negative qualities
  - voraciously, blarney, rashness, laboring, ravage, destined, reproaches, desolation, solicitations, inhumanly, suitor

Sample from 100 words whose change curves correlate best with the mean – SDM

- Endeavors and objectives of the individual, positive qualities
  - plentifully, shelter, content, devoutly, animated, charms, escape, emancipate, spiritual, beautiful, agreeably, fortune, deliciously, convenience, thrilling, diligent, familiar, fancying, lofty, souls, blessedness, lustre
- Troubles that the individual may encounter, negative qualities
  - voraciously, exploit, rashness, toil, desolation, hesitation, solicitations, strive, suitors, reckless, troubled

The “me generation”

The “Me” Decade and the Third Great Awakening

“. . . The new alchemical dream is: changing one’s personality—remaking, remodeling, elevating, and polishing one’s very *self* . . . and observing, studying, and doting on it. (Me!) . . .”

By Tom Wolfe

2 Comments

Tom Wolfe at the White House

HEALTHY LIVING 03/21/2017 04:52 pm ET | **Updated** 3 days ago

Are Baby Boomers A ‘Generation Of Sociopaths’?

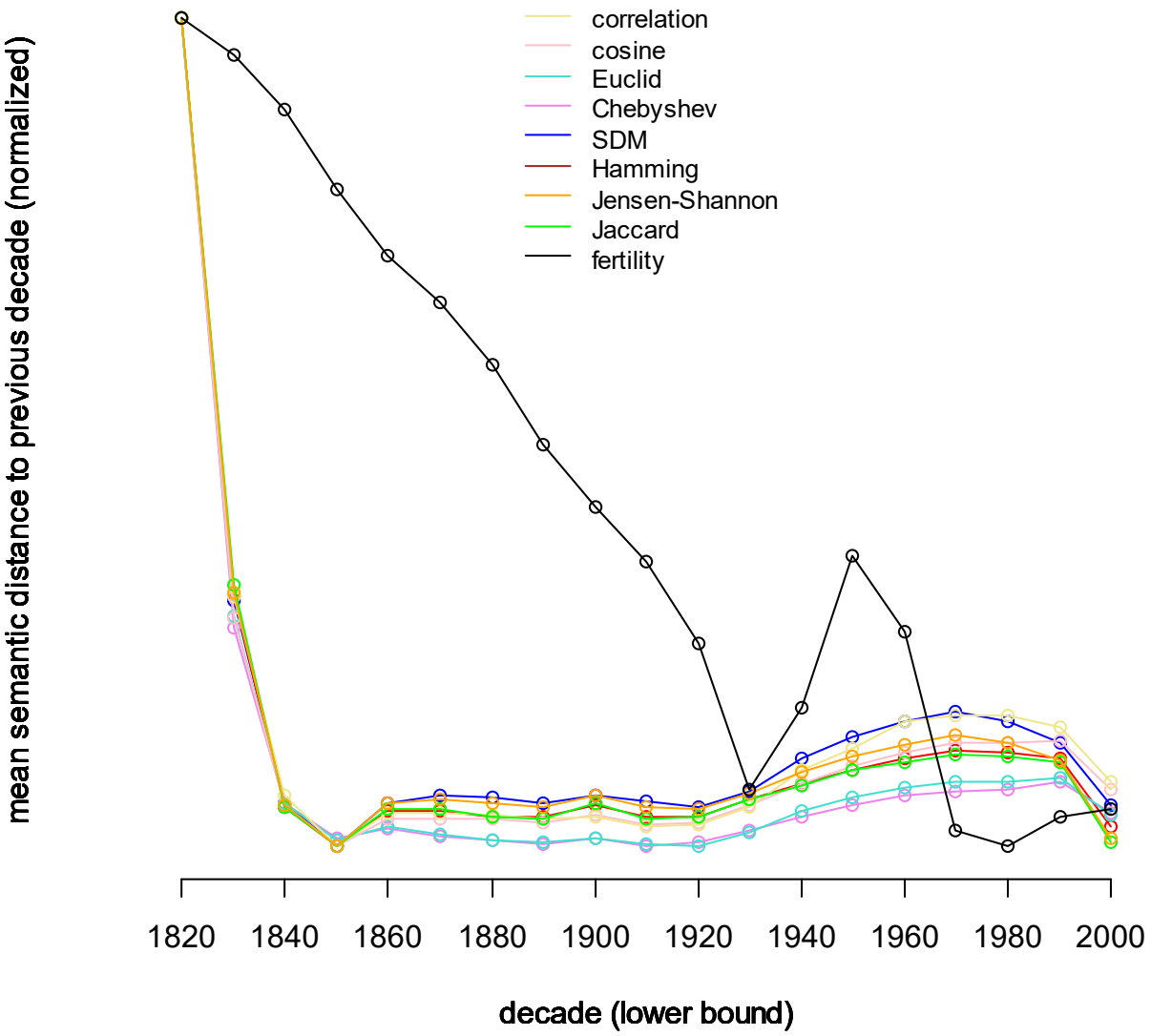
A controversial new book argues boomers are beset with egotism, impulsivity and a shocking lack of empathy — and they’re leaving the world a worse place.

By Carolyn Gregoire

MARK HUNT

Long before millennials were dubbed the “Me Generation,” journalist Tom Wolfe used the label to describe the young baby boomers coming of age in the mid-1970s, a time of heightened focus on the self and personal development.

http://www.huffingtonpost.com/entry/baby-boomers-sociopaths_us_58b9a358e4b0d2821b4dd797


correlation
cosine
Euclid
Chebyshev
SDM
Hamming
Jensen-Shannon
Jaccard
fertility
mean semantic distance to previous decade (normalized)
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
decade (lower bound)

Conclusions

- Different vector semantic measures are sensitive in different ways to frequencies
  - needs to be studied in more detail
  - worth considering there may not be a best measure—they may complement each other for some purposes
  - more recent methods worth exploring
- Average semantic change in the core lexicon seems to driven by words relating to individualism
  - Can this somehow be established statistically?

References

- Jurafsky, Daniel and James H. Martin. 2016. Speech and Language Processing. Chapter 15: Vector semantics. Draft of November 7, 2016.
- Lee, Lillian. 1999. Measures of distributional similarity. In ACL-99, pp. 25–32.
- Teknomo, Kardi: similarity measurement. https://people.revoledu.com/kardi/tutorial/Similarity/index.html