ETH zürich

How group identity and human-computer interaction shape online communication

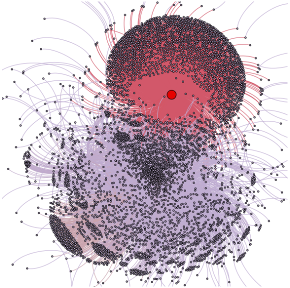
David García

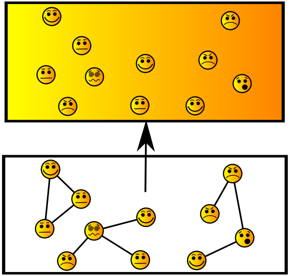
with M. Strohmaier, C. Wagner, E. Graells-Garrido, F. Menczer

IPAM
www.sg.ethz.ch

Chair of Systems Design at ETH Zurich

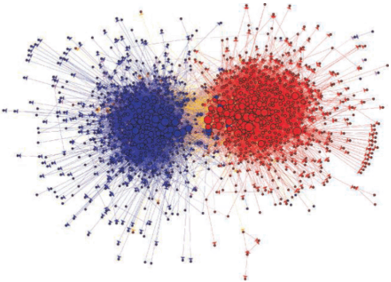
- Data Driven Modeling & Computational Social Science





- Retrieval and Analysis of Digital traces
  - **Networks**: Social networks, communication, reputation, resilience
  - **Text**: Public messages, product reviews, sentiment analysis
  - **Dynamics**: Time series analysis, complex systems, collective emotions

Computational Social Science



Computational Social Science

Quantitative testing theories from the social sciences at unprecedented breadth and depth and scale

(Lazer et. al. Science, 2009)

- Quantitative, empirical vs previous qualitative and theoretical work
- Not data-driven descriptive studies: **research between disciplines**
- Towards computational models to understand human behavior

Related but not the same as Big Data, Data Science, Web Science, Digital Sociology, Human-Computer Interaction, Behavioral Science...

Digital, Computerized, and Generative

The *Computational* in Computational Social Science means:

Digital

Based on large datasets of human behavior, either produced by the Web and social media, or on digital databases of culture and History

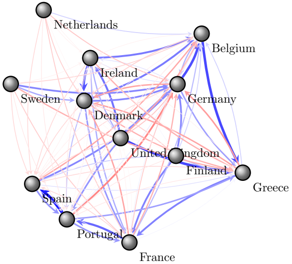
Computerized

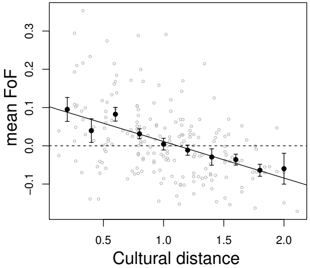
The quantitative analysis of data in an automated, tractable, repeatable, and extensible fashion

Generative

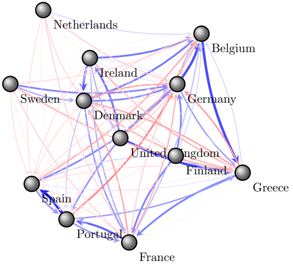
Application of data and results to design of agent-based models that explain observed social phenomena and motivate interventions

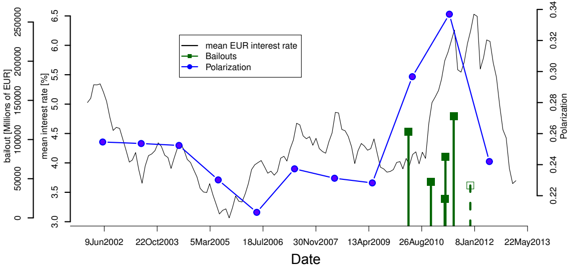
Digital traces of cultural distance: Eurovision

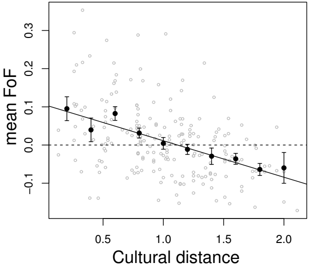




Digital traces of cultural distance: Eurovision

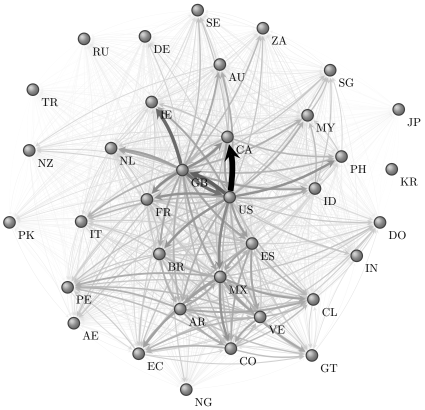


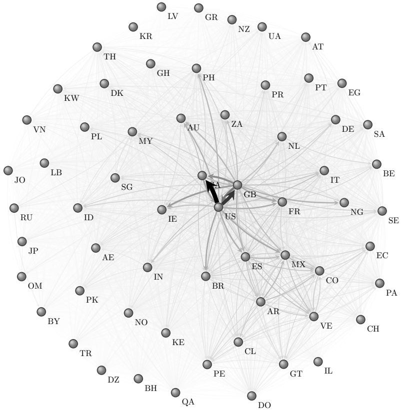




Measuring cultural dynamics through the Eurovision song contest. D. Garcia, D. Tanase. *Advances in Complex Systems.* 2013

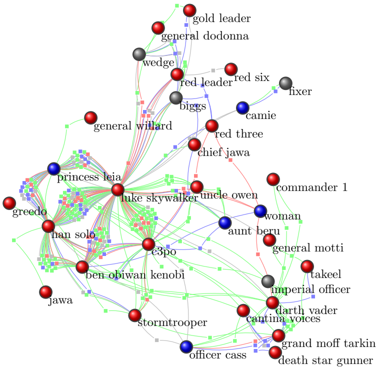
Digital traces of cultural distance: Twitter TTs

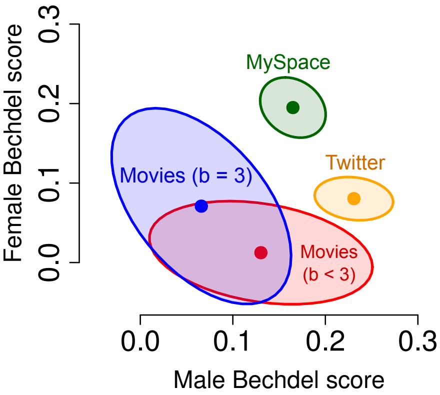




Quantifying the Economic and Cultural Biases of Social Media through Trending Topics. J. M. Carrascosa, R. Cuevas , R. Gonzalez, A. Azcorra, D. Garcia. *PLoS ONE.* 2015

The Bechdel test of social media

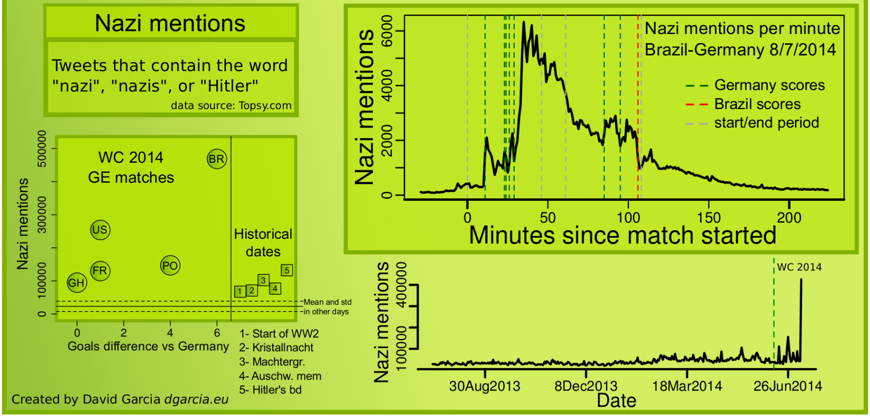




Gender Asymmetries in Reality and Fiction: The Bechdel Test of Social Media D. Garcia, I. Weber, V. Garimella. *ICWSM* (2014)

Godwin's law in the World Cup 2014

As the German national team scores goals in a soccer match,
the probability of a comparison involving Nazis or Hitler approaches 1




Godwin's law in the World Cup 2014
As the German national team scores goals in a soccer match,
the probability of a comparison involving Nazis or Hitler approaches 1
Nazi mentions
Tweets tha
"nazi", "na
WC 2
GE mat
Nazi mentions per minute
7/2014
scores
es
eriod
LAST SLIDE BASED
ON TWITTER DATA
Nazi mentions
0
100000
300000
500000
US
FR
GH
Mean and std
in other days
0
2
4
6
Goals difference vs Germany
1- Start of WW2
2- Kristallnacht
3- Machtergr.
4- Auschw. mem
5- Hitler's bd
Nazi ment
100000
30Aug2013
8Dec2013
18Mar2014
26Jun2014
Date
C 2014
Created by David Garcia dgarcia.eu

Outline

1. The Linguistic Intergroup Bias
   - Group identity in language
   - Wikipedia study
   - The gig economy study

2. The QWERTY effect
   - Keyboards in communication
   - Decoding Study
   - Encoding Study

Group identity in communication

The Linguistic Intergroup Bias (Maass et. al. 1989)

People encode and communicate desirable in-group and undesirable out-group behaviors more abstractly than undesirable in-group and desirable out-group behaviors.

Group identity in communication

The Linguistic Intergroup Bias (Maass et. al. 1989)

People encode and communicate desirable in-group and undesirable out-group behaviors more abstractly than undesirable in-group and desirable out-group behaviors.

- Barcelona newspaper:
  "Messi *committed* a *foul* in the match"

Group identity in communication

The Linguistic Intergroup Bias (Maass et. al. 1989)

People encode and communicate desirable in-group and undesirable out-group behaviors more abstractly than undesirable in-group and desirable out-group behaviors.

- Barcelona newspaper:
  "Messi *committed* a *foul* in the match"
- Madrid newspaper:
  "Messi: the *violent and aggressive* player"

Group identity in communication

The Linguistic Intergroup Bias (Maass et. al. 1989)

People encode and communicate desirable in-group and undesirable out-group behaviors more abstractly than undesirable in-group and desirable out-group behaviors.

- Barcelona newspaper:
  "Messi *committed* a *foul* in the match"
- Madrid newspaper:
  "Messi: the *violent and aggressive* player"
- National newspaper:
  "Riots after R. Madrid - F.C. Barcelona"

The digital traces of LIB

Analysis of LIB (Otterbacher, 2015)

1. identification of sentiment through the subjectivity clues lexicon

   `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`

   Word subjectivity annotations

   - 2718 positive
   - 4912 negative
   - 570 neutral
   - 21 both
2. detection of abstract language through POS tagging
   - adjectives are abstract
   - verbs can be abstract or concrete

Linguistic Intergroup Bias in Wikipedia

LIB in Wikipedia?

- With respect to gender: The vast majority Wikipedia editors are male
- LIB for biographies depending on gender?
- Should not exist: Wikipedia neutrality/plurarism regulations
- Should not exist: Readers of articles are undetermined

Data Summary

- Wikipedia biographies from DBpedia 2014
- Selection of biographies with at least 250 words
- Gender from (Bamman and Smith, 2014) annotations
- $\sim$ 50K biographies

Detecting abstract language

Tendency to express positive traits in an abstract manner:

Postive abstract ratio

$$r_+ = \frac{\# \quad positive \quad adjectives}{\# \quad positive \quad terms}$$

Tendency to express negative traits in an abstract manner:

Negative abstract ratio

$$r_- = \frac{\# \quad negative \quad adjectives}{\# \quad negative \quad terms}$$

LIB in Wikipedia

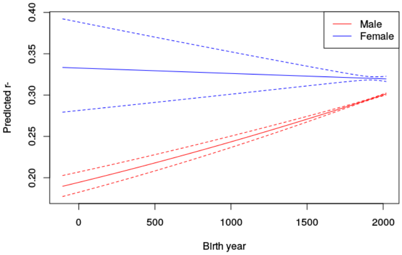
	% in men	% in women	χ^2	w	% change
Abstract positive	27.96	25.53	933.7***	0.04	8.69
Abstract negative	13.47	13.69	6.26**	0.005	−1.62

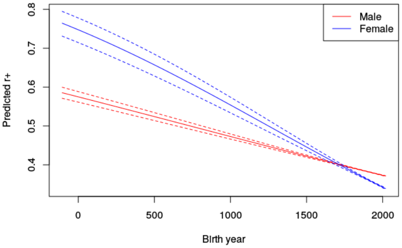
Comparison of the ratios of abstract terms among positive and negative terms for men and women. Slightly more abstract terms are used for positive aspects in men's biographies, while slightly more abstract terms are used for negative aspects in women's biographies. ***: $p<0.001$, **: $p<0.01$.

LIB in Wikipedia

	% in men	% in women	χ^2	w	% change
Abstract positive	27.96	25.53	933.7***	0.04	8.69
Abstract negative	13.47	13.69	6.26**	0.005	−1.62

Comparison of the ratios of abstract terms among positive and negative terms for men and women. Slightly more abstract terms are used for positive aspects in men's biographies, while slightly more abstract terms are used for negative aspects in women's biographies. ***: $p<0.001$, **: $p<0.01$.





Resources

Wagner et al. *EPJ Data Science* (2016) 5:5
DOI 10.1140/epjds/s13688-016-0066-4
EPJ.org
EPJ Data Science
a SpringerOpen Journal
REGULAR ARTICLE Open Access
CrossMark

Women through the glass ceiling: gender asymmetries in Wikipedia

Claudia Wagner[1,2*], Eduardo Graells-Garrido[3], David Garcia[4] and Filippo Menczer[5]

`https://github.com/dgarcia-eu/LIB_Tutorial`

Linguistic Intergroup Bias in the gig economy

1 The Linguistic Intergroup Bias
- Group identity in language
- Wikipedia study
- The gig economy study

2 The QWERTY effect
- Keyboards in communication
- Decoding Study
- Encoding Study

LIB in the gig economy?

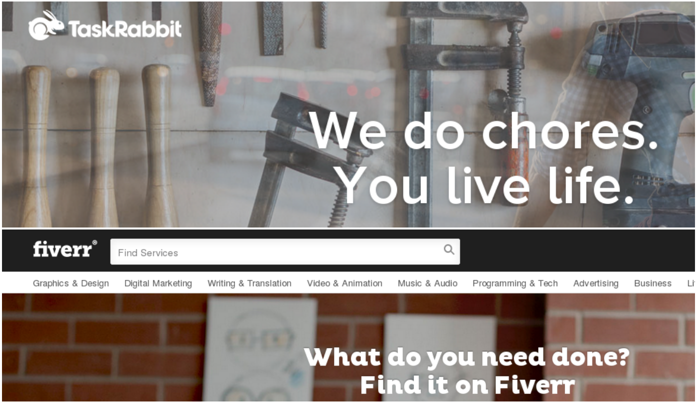
The gig economy

A gig economy is an environment in which temporary positions are common and organizations contract with independent workers for short-term engagements. (techtarget.com)

- Job stability depends on customer reviews and ratings
- Minimal moderation of feedback
- Could there be LIB in the reviews given by gig economy customers?

collaborators: A. Hannak, C. Wagner, M. Strohmaier, C. Wilson, A. Mislove

Data on the gig economy



Task Rabbit dataset

- ∼ 3K users (53% male)
- ∼ 54K reviews

Fiverr dataset

- ∼ 9K workers (63% male)
- ∼ 136K reviews

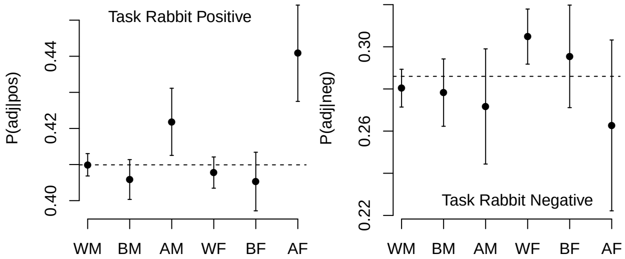
Sentiment-bearing word abstraction

Word-level model

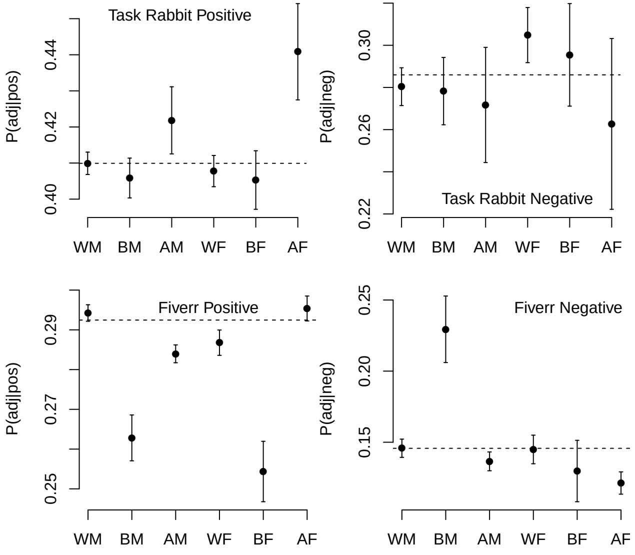
Given that a positive word is uttered in a review for a worker of certain race and gender, what is the probability that such word is an adjective?

- Logistic regression model over words with gender, race, and their interaction as explanatory variables
- Controls: average rating of the worker, amount of gigs of the worker, experience, etc
- Effect visualization of probabilities depending on combinations of race and gender

LIB in Task Rabbit and Fiverr



LIB in Task Rabbit and Fiverr



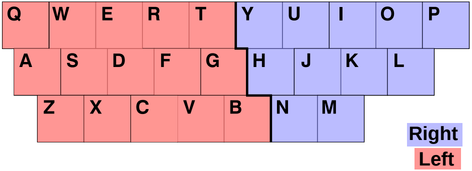
Discussions and caveats

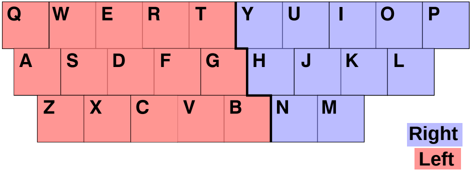
1. Different results for Task Rabbit and Fiverr
   - Effect of gender and race ratios?
2. No information on identity of the writers of texts
   - LIB or general discrimination?
3. Signals of style
   - It is not what is said, it is how it is said
4. Wikipedia effects
   - Does time weaken the LIB?
   - Can we observe the LIB across Wikipedia languages?
5. More advanced models?
   - Including topics
   - Recovering LIB from representations
   - Predictive formulations

The QWERTY effect

What were the body organs that you used the last time that you communicated with someone?

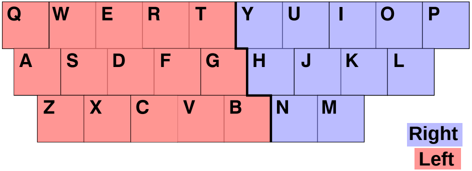
What were the body organs that you used the last time that you communicated with someone?


Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
Right
Left



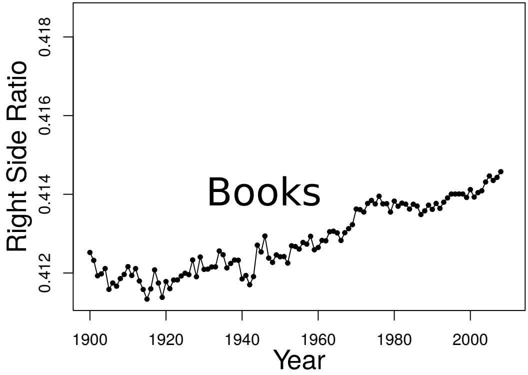
The Right Side Ratio

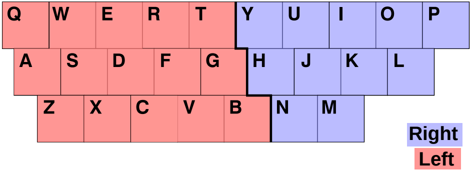
$$RSR = \frac{R}{R+L}$$



The Right Side Ratio

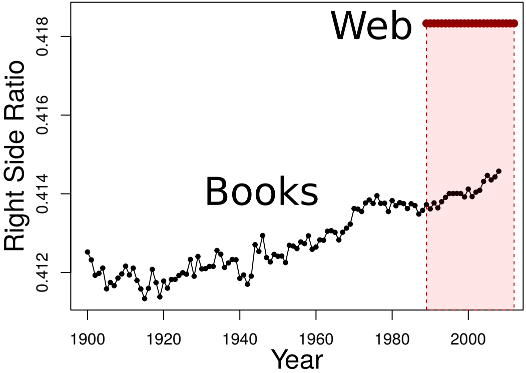
$$RSR = \frac{R}{R+L}$$





The Right Side Ratio

$$RSR = \frac{R}{R+L}$$



Psychon Bull Rev (2012) 19:499–504
DOI 10.3758/s13423-012-0229-7

BRIEF REPORT

The QWERTY Effect: How typing shapes the meanings of words.

Kyle Jasmin · Daniel Casasanto

The QWERTY effect hypothesis

On average, words typed with more letters from the right side of the keyboard are more positive in meaning than words typed with more letters from the left.

Psychon Bull Rev (2012) 19:499–504
DOI 10.3758/s13423-012-0229-7

BRIEF REPORT

The QWERTY Effect: How typing shapes the meanings of words.

Kyle Jasmin · Daniel Casasanto

The QWERTY effect hypothesis

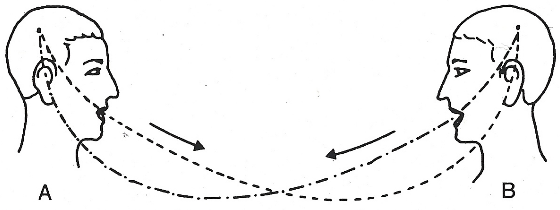
On average, words typed with more letters from the right side of the keyboard are more positive in meaning than words typed with more letters from the left.

Previous evidence

- Evidence in word ratings for English, Spanish, Dutch, and German
- Stronger for post-QWERTY neologisms and present for pseudowords

The research gap

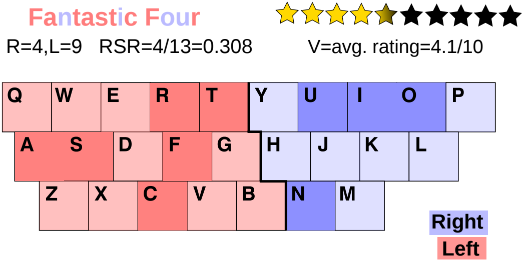
1. Previous evidence from small scale subjective ratings experiments
   - Limitations: design issues, rater disagreement, external validity
   - Observational evidence on baby names is inconclusive
   - Can we observe the QWERTY effect **on the Web**?

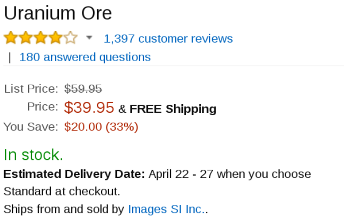


   - *Decoding:* interpreting the meaning of words (reading)
   - *Encoding:* translating meaning into words (writing)
   - Previous evidence only on *decoding*
2. Can we observe the QWERTY effect in *encoding* as well?

Decoding Study

Decoding study design





- Valence *V*: positivity of crowdsourced evaluation (average rating, percentage of likes...)
- Right-Side Ratio *RSR*: of the title or name of what is evaluated
- We additionally measure contextual and linguistic controls (word length, frequency, popularity)

Decoding study datasets

Dataset	N elements	$\langle V \rangle$	scale	$\langle RSR \rangle$	source
Amazon	4,257,624	3.86	1-5	0.4176	(McAuley, 2015)
Yelp	56,103	3.66	1-5	0.4056	Yelp Challenge
Epinions	223,880	3.89	1-5	0.4174	(Tanase, 2013)
Dooyoo	112,698	3.89	1-5	0.4184	(Tanase, 2013)
IMDB	327,608	6.30	1-10	0.425	OMDB
Rotten Tomatoes	80,756	3.04	1-5	0.4233	OMDB
MovieLens	29,505	3.11	1-5	0.4245	(Harper, 2015)
BookCrossing	149,804	7.42	1-10	0.4164	(Ziegler, 2005)
Youtube	3,292,153	0.94	0/1	0.4294	(Abisheva, 2014)
Redtube	351,677	0.70	0/1	0.4225	new
Pornhub	333,967	0.83	0/1	0.4264	new

Statistical analysis methods

1. Linear effect analysis:

$$V = a + b * RSR + \epsilon$$

- Hypothesis: Right-side coefficient $b > 0$
- 5 Methods: OLS, MM-robust, bootstrapping, permutation, Spearman

2. Residualized regression to control for possible confounding factors

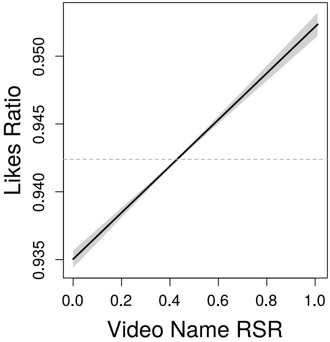
$$V = \sum_i c_i * X_i + V_r \qquad V_r = a' + b' * RSR + \epsilon'$$

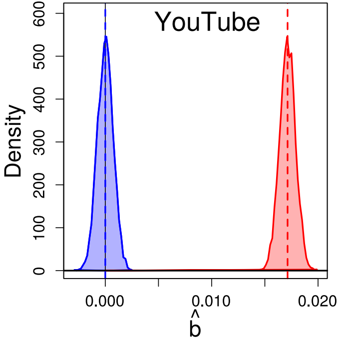
- 2.1 Residualized OLS test
- 2.2 When data very large: stratified regression on each control variable

YouTube Video Likes

Dataset	N elements	$\langle V \rangle$	scale	$\langle RSR \rangle$
Youtube (Abisheva et al, 2014)	3,292,153	0.94	0/1	0.4294

controls: name len, avg word f, avg letter f, nwords, views, comms, date, N_r



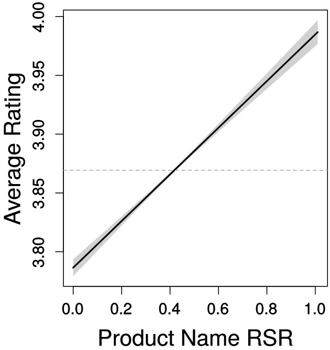


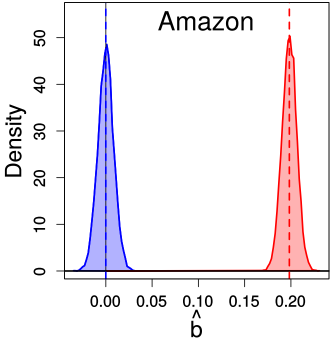
OLS $\hat{b}$	MM $\hat{b}$	Resid. $\hat{b}$	Bootstrap	Permutation	Spearman
0.0171	**0.0007**	**0.0109**	$p < 0.05$	$p < 0.05$	$p < 0.05$

Amazon Product Ratings

Dataset	N elements	$\langle V \rangle$	scale	$\langle RSR \rangle$
Amazon (McAuley et al, 2015)	4,257,624	3.86	1-5	0.4176

controls: name len, avg word freq, avg letter freq, nwords sales rank, price, N_r



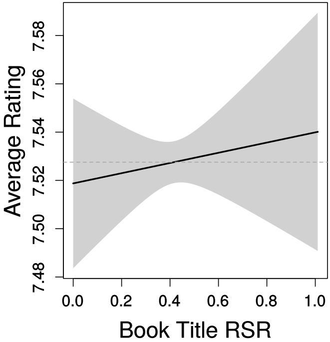


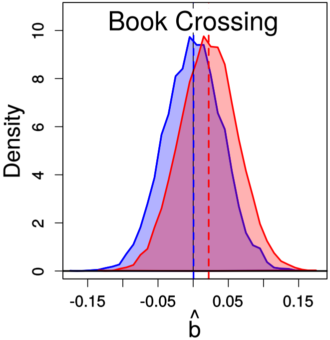
OLS $\hat{b}$	MM $\hat{b}$	Resid. $\hat{b}$	Bootstrap	Permutation	Spearman
0.1984	**0.1384**	**0.0348**	$p < 0.05$	$p < 0.05$	$p < 0.05$

Book Crossing Book Ratings

Dataset	N elements	$\langle V \rangle$	scale	$\langle RSR \rangle$
BookCrossing (Ziegler et al, 2005)	149,804	7.42	1-10	0.4164

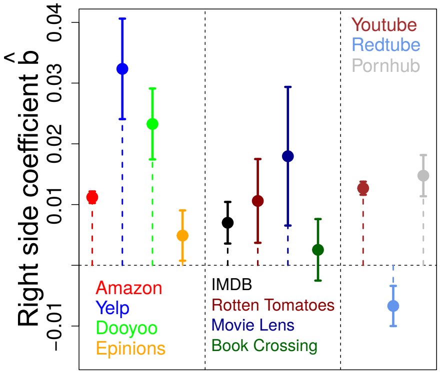
controls: name len, avg word freq, avg letter freq, nwords, N_r





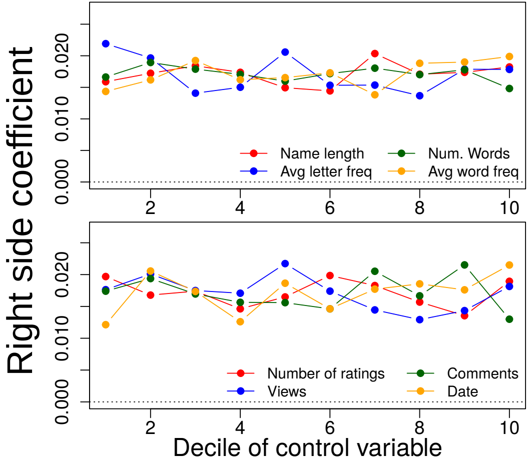
OLS $\hat{b}$	MM $\hat{b}$	Resid. $\hat{b}$	Bootstrap	Permutation	Spearman
0.0414	*0.0516*	*-0.0408*	$p > 0.1$	$p > 0.1$	$p > 0.1$

Decoding results summary



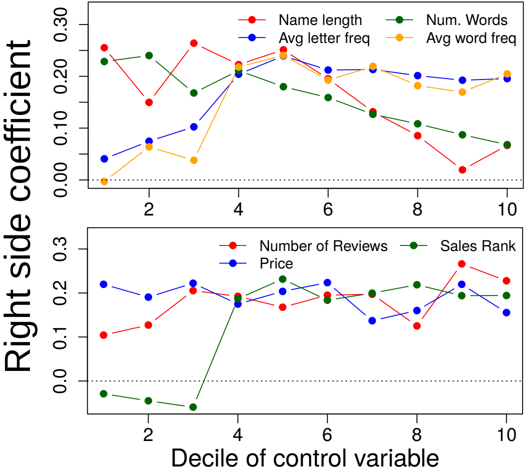
- Renormalized estimate of the Right side coefficient $\widehat{b}$
- Estimate is positive and significant for 9 out of 11 datasets
- Youtube videos with more right letters in the title get more likes
- Products with more right letters in the name have higher ratings
- Movies with more right letters in the title get better ratings

Understanding the effect of context: Youtube



- Regression analysis over subsets selected by control variable
- **YouTube:** Effect still present for changing controls

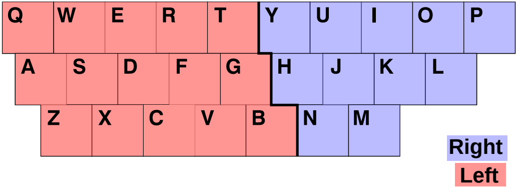
Understanding the effect of context: Amazon



- Regression analysis over subsets selected by control variable
- **Amazon:** Effect disappears for high sales and infrequent language

Encoding Study

Encoding study design and methods



star rating ➤ V

qwer yuio ta ds jkl fz xvw
qe nm rre nmafdsf as vz
cjhjhx vfdwkkq erfdashl
vdzcv zhnndfre wqe woy
ioihok freaz vz

Review text ➤ R, L

1. $V = a_l + b_l * Length$
2. $V = a_{RL} + b_R * R + b_L * L + b_{RL} * R * L$

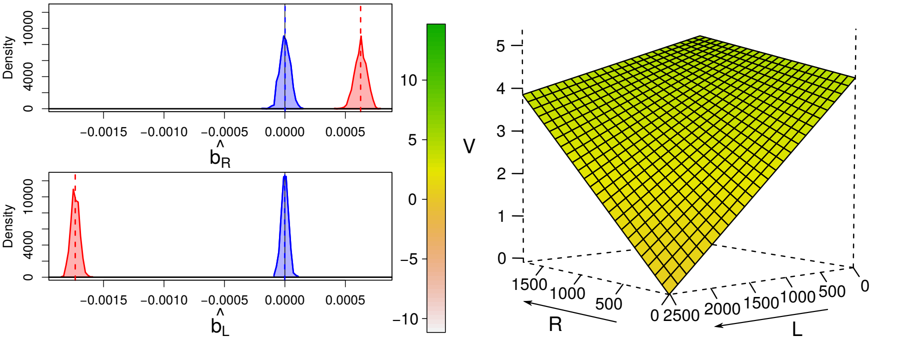
- MM-robust ΔR^2_{adj}
- bootstrapping and permutation t.
- Hypotheses: $b_R > 0$ and $b_L < 0$

Datasets:

Website	N reviews	$\langle r \rangle$
Amazon	971,026	4.0582
Yelp	1,554,163	3.7412
Dooyoo	523,997	4.0258
Epinions	101,595	3.9768

Yelp Business Reviews

Dataset	N reviews	$\langle r \rangle$	scale
Yelp	1,554,163	3.7412	1-5



Length model		RL model				
a_l	b_l	a_{RL}	b_R	b_L	b_{RL}	ΔR^2_{adj}
4.172784	-0.000458	4.263156	**0.000625**	**-0.001733**	10^{-7}	8.2%

Encoding study results

Length Model:

Dataset	a_l	b_l
Yelp	4.172784	-0.000458
Amazon	4.096829	-0.000123
Dooyoo	4.255313	0.000014
Epinions	4.274520	0.000007

RL Model:

Dataset	a_{RL}	b_R	b_L	b_{RL}	ΔR^2_{adj}
Yelp	4.263156	**0.000625**	-**0.001733**	10^{-7}	8.2%
Amazon	4.110348	**0.00086**	-**0.000940**	10^{-8}	40.4%
Dooyoo	4.256541	**0.000169**	-**0.000101**	0.00	14.5%
Epinions	4.278481	*0.000048*	*-0.000027*	0.00	-5.1%

Discussion: Interpretations

1- The theory
Using the keyboard influences word meanings

Discussion: Interpretations

1- The theory

Using the keyboard influences word meanings

2- The design

Was the QWERTY design influenced by meanings?

Discussion: Interpretations

1- The theory

Using the keyboard influences word meanings

2- The design

Was the QWERTY design influenced by meanings?

3- The confound

Is this an effect of phonetic/linguistic properties?

Discussion: Interpretations

1- The theory

Using the keyboard influences word meanings

2- The design

Was the QWERTY design influenced by meanings?

3- The confound

Is this an effect of phonetic/linguistic properties?

4- The unexpected

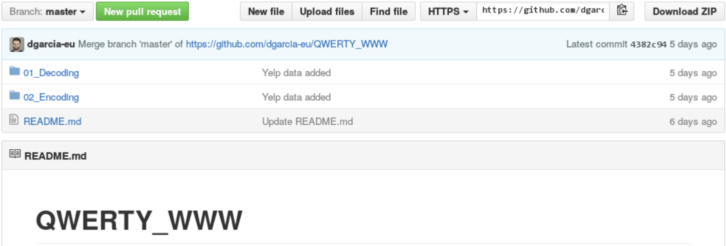
Is there an army of one-handed spambots?

Discussion: Caveats

1. Beware $N = all$
   - Representativity warning: Big data but narrow media
   - No selection of data: Additional clusters and correlations
2. We only saw the traces of the QWERTY effect
   - Lack of control opposed to experimental studies
   - Missing questions: Other languages, effect of handedness...
3. We only test the QWERTY effect hypothesis
   - No evidence that we can predict meaning from the RSR
   - No evidence that we can change evaluations throught the RSR
4. Small effects might not be so small
   - Interpreting effect sizes in language is not trivial

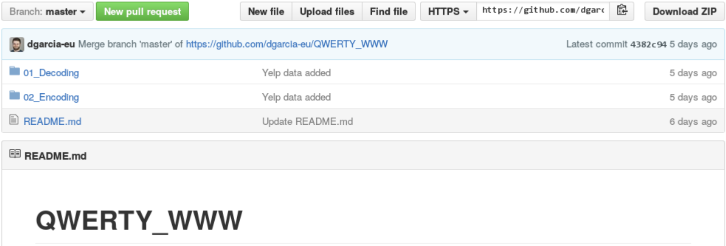
Conclusion

- We found evidence of the QWERTY effect when encoding and decoding text from the Web in a wide variety of media
- We found some limiting factors and counterexamples
- Data and codes: `https://github.com/dgarcia-eu/QWERTY_WWW`



Conclusion

- We found evidence of the QWERTY effect when encoding and decoding text from the Web in a wide variety of media
- We found some limiting factors and counterexamples
- Data and codes: `https://github.com/dgarcia-eu/QWERTY_WWW`



Just a recommendation

If you doubt between two names: Choose the *right* one!

Thanks for listening!

More in: `dgarcia.eu` and Twitter: @dgarcia_eu