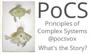
Linguistic Pollyanna Principle: The positivity bias of language

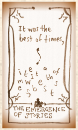
Last updated: 2023/08/26, 09:18:43 EDT

Principles of Complex Systems, Vols. 1, 2, & 3D CSYS/MATH 6701, 6713, & a pretend number, 2023–2024| @pocsvox

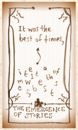
Prof. Peter Sheridan Dodds | @peterdodds

Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont





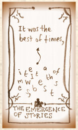
These slides are brought to you by:



These slides are also brought to you by:

Special Guest Executive Producer

On Instagram at pratchett_the_cat



Outline

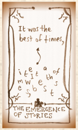
Pollyanna Principle

English is happy

10 languages

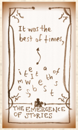
Extras
- Corpora
- Text parsing
- Corpus generation

References




Human language reveals a universal positivity bias

“Human language reveals a universal positivity bias”
Dodds et al.,
Proc. Natl. Acad. Sci., **112**, 2389–2394, 2015. [2]



Who are we?

- Stories we tell about how we should/could/must behave vary enormously.
- Jainism to Rand's Objectivism.

Basic observations:

- Language is our great social technology.
- And we convey stories through language.

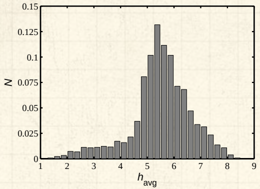
Basic question:

- What's the distribution of emotional content of the atoms of language?

Data we've generated:

- English plus nine other languages.
- Key: incorporate word usage frequency (= size).

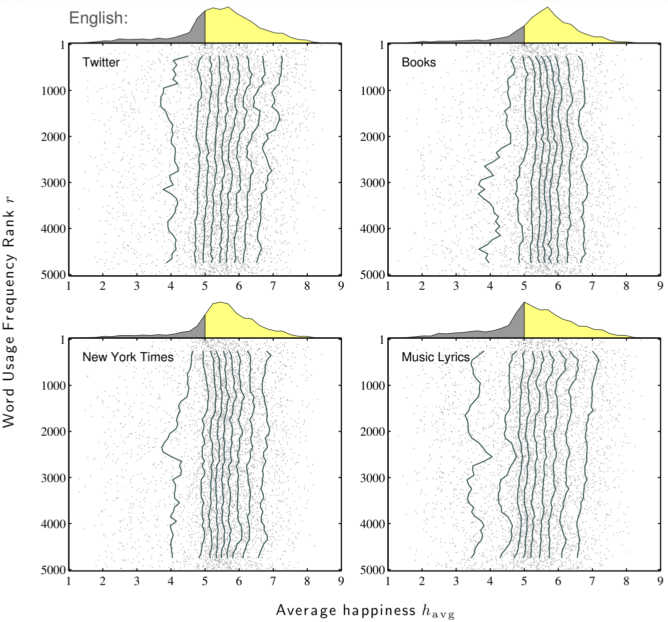
English's scale-invariant, positive bias: [8]

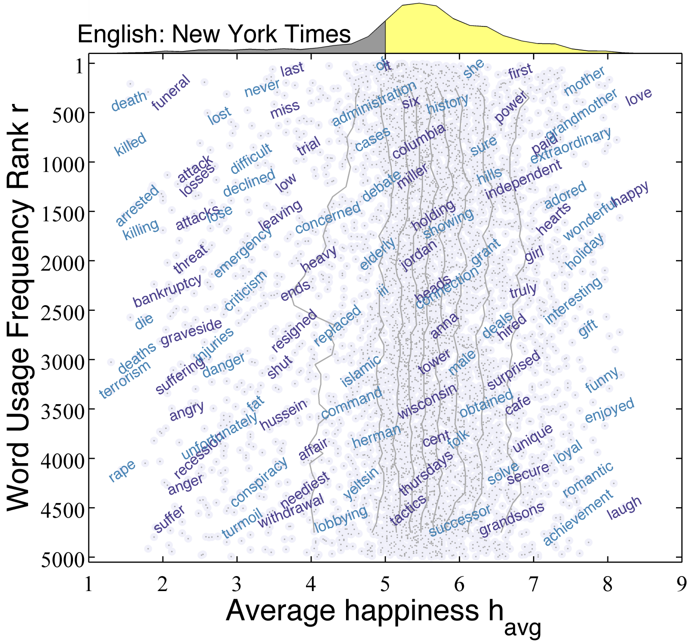


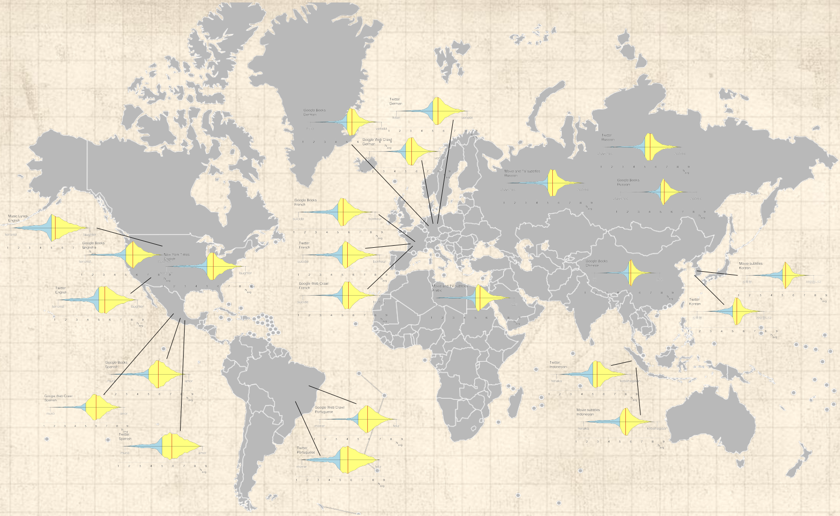
- Social organism story manifested in language.
- Pollyanna Hypothesis: Interactions are predominantly positive

- Positive anchor of concepts: Unhappy but not unsad.
- Many ways for things to go wrong: "All happy families are alike; each unhappy family is unhappy in its own way."
- Guns, Germs, and Steel [1] invokes the Anna Karenina Principle
- But: must account for frequency of word usage ...

Jellyfish plots:




English: New York Times
Word Usage Frequency Rank r
Average happiness h_{avg}
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7
8
9
death
funeral
never
last
of
it
she
first
mother
love
lost
miss
administration
six
history
power
grandmother
killed
trial
cases
columbia
sure
paid
extraordinary
attack
losses
difficult
declined
low
debate
miller
hills
independent
arrested
killing
attacks
lose
leaving
concerned
holding
showing
adored
hearts
wonderful
happy
threat
emergency
heavy
elderly
jordan
grant
girl
holiday
bankruptcy
criticism
ends
iii
heads
connection
truly
interesting
die
graveside
resigned
replaced
anna
deals
hired
gift
deaths
terrorism
suffering
injuries
danger
shut
islamic
tower
male
surprised
funny
angry
unfortunately
fat
hussein
command
wisconsin
obtained
cafe
enjoyed
rape
recession
anger
conspiracy
affair
herman
cent
folk
unique
loyal
neediest
yeltsin
thursdays
solve
secure
romantic
suffer
turmoil
withdrawal
lobbying
tactics
successor
grandsons
achievement
laugh



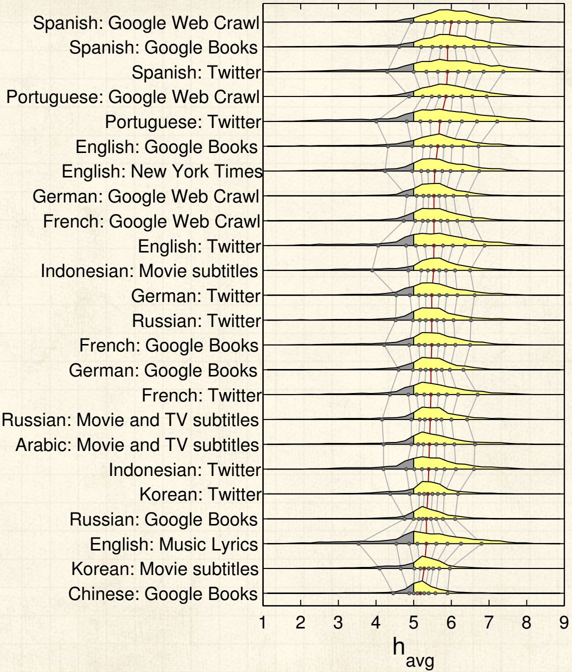
Dodds/Tivnan/Danforth et al.,
Proc. Natl. Acad. Sci. 2015,
"Human language reveals a universal positivity bias." [2]
Global press including National Geographic
Top 100 altmetric article, 2015

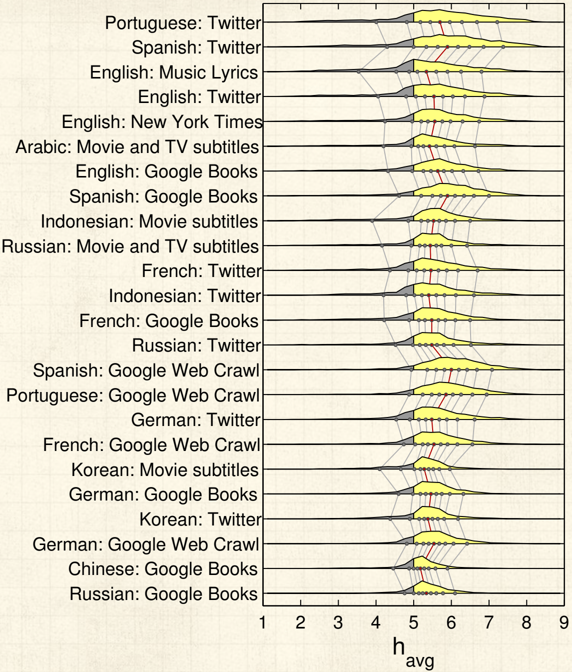
Good buzz according to Altmetric ...(report is no longer findable):

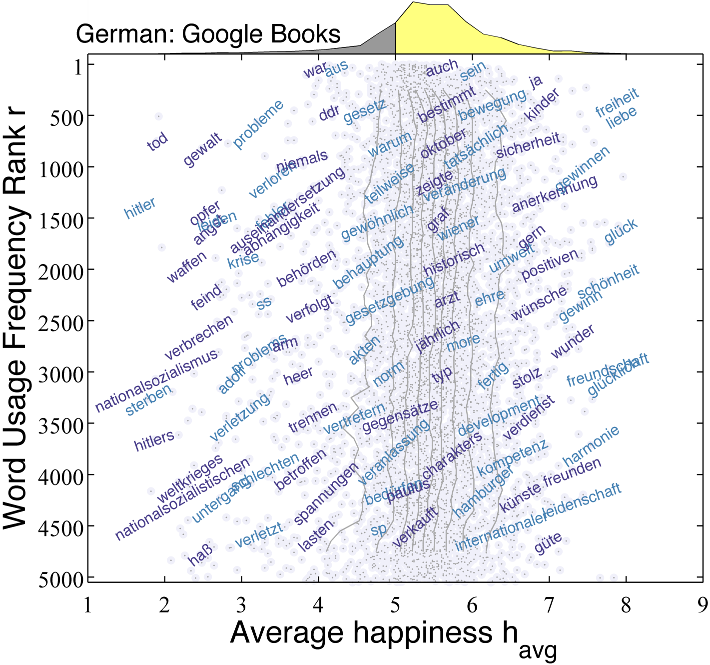
As of May 7, 2015:

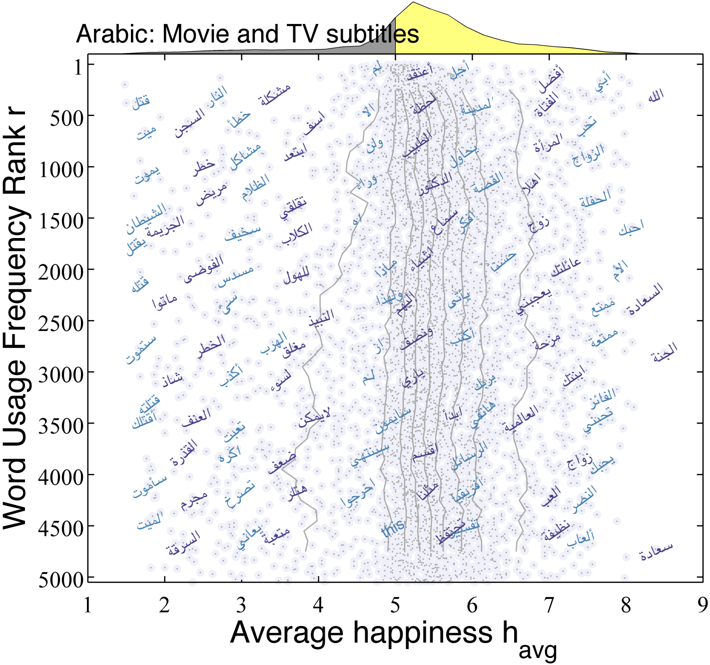
- Altmetric Score: 772.
- Ranked 3rd out of 933 articles published in PNAS surrounding 12 weeks.
- Ranked 24nd out of 34,050 articles in PNAS all time. (Mean score 13.5.)
- Ranked 60th out of all 109,841 tracked articles published in surrounding 12 weeks.
- Ranked 459th out of 3,724,005 tracked articles all time.

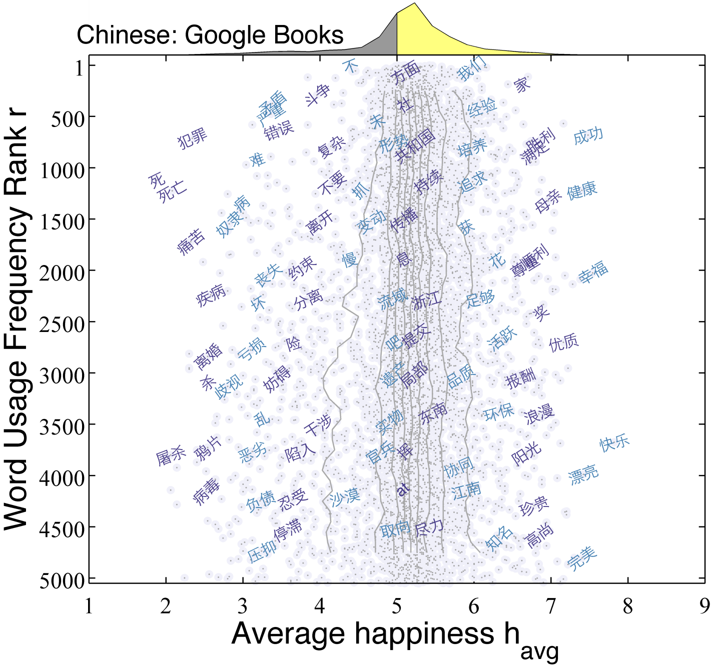
This doesn't mean it's a good article ... but it is.


Spanish: Google Web Crawl
Spanish: Google Books
Spanish: Twitter
Portuguese: Google Web Crawl
Portuguese: Twitter
English: Google Books
English: New York Times
German: Google Web Crawl
French: Google Web Crawl
English: Twitter
Indonesian: Movie subtitles
German: Twitter
Russian: Twitter
French: Google Books
German: Google Books
French: Twitter
Russian: Movie and TV subtitles
Arabic: Movie and TV subtitles
Indonesian: Twitter
Korean: Twitter
Russian: Google Books
English: Music Lyrics
Korean: Movie subtitles
Chinese: Google Books
1 2 3 4 5 6 7 8 9
h_{avg}

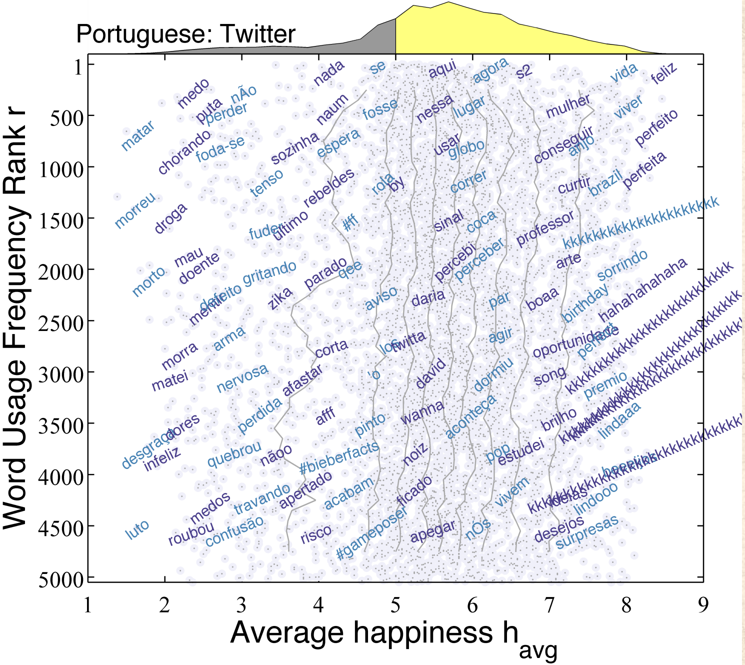

Portuguese: Twitter
Spanish: Twitter
English: Music Lyrics
English: Twitter
English: New York Times
Arabic: Movie and TV subtitles
English: Google Books
Spanish: Google Books
Indonesian: Movie subtitles
Russian: Movie and TV subtitles
French: Twitter
Indonesian: Twitter
French: Google Books
Russian: Twitter
Spanish: Google Web Crawl
Portuguese: Google Web Crawl
German: Twitter
French: Google Web Crawl
Korean: Movie subtitles
German: Google Books
Korean: Twitter
German: Google Web Crawl
Chinese: Google Books
Russian: Google Books
1
2
3
4
5
6
7
8
9
h_{avg}

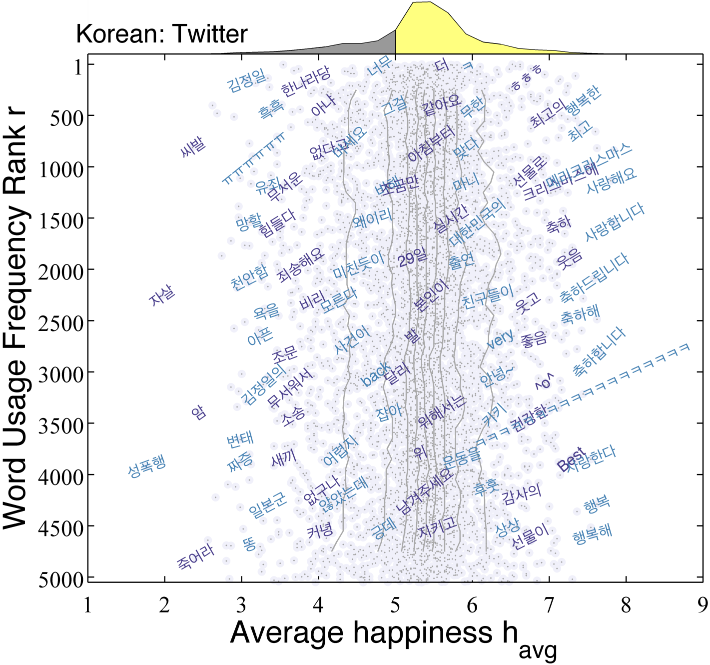

German: Google Books
Word Usage Frequency Rank r
Average happiness h_{avg}
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1 2 3 4 5 6 7 8 9

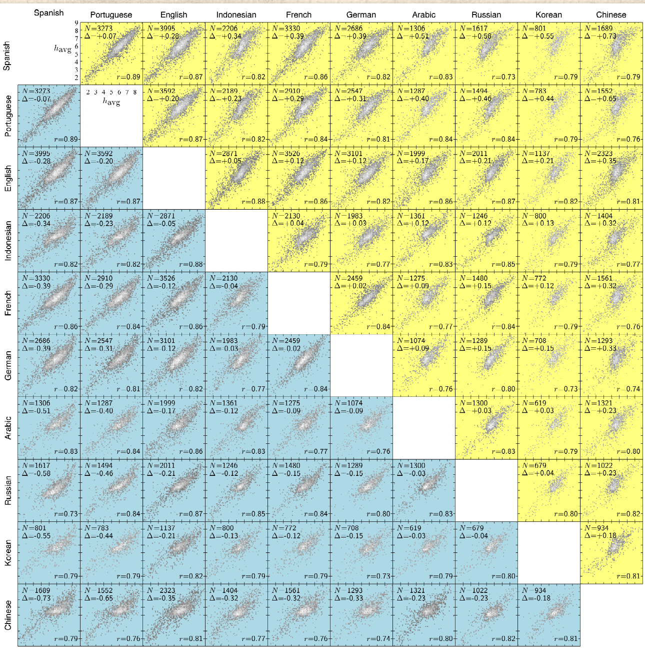

Arabic: Movie and TV subtitles
Word Usage Frequency Rank r
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7
8
9
Average happiness h_{avg}
this



Chinese: Google Books
Word Usage Frequency Rank r
Average happiness h_{avg}
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1 2 3 4 5 6 7 8 9
不
方面
我们
家
矛盾
严重
斗争
社
经验
犯罪
错误
末
复杂
形势
共和国
培养
胜利
满足
成功
难
死
死亡
不要
抓
持续
追求
母亲
健康
奴隶病
痛苦
离开
变动
传播
获
慢
丧失
约束
息
花
尊重
幸福
疾病
坏
分离
流域
浙江
足够
奖
离婚
亏损
险
吧
提交
活跃
优质
杀
歧视
妨碍
遗产
局部
品质
报酬
乱
干涉
实物
东南
环保
浪漫
屠杀
鸦片
恶劣
陷入
官兵
择
阳光
快乐
协同
病毒
负债
忍受
沙漠
at
江南
漂亮
珍贵
压抑
停滞
取向
尽力
知名
高尚
完美


Portuguese: Twitter
Word Usage Frequency Rank r
Average happiness h_{avg}
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7
8
9

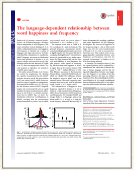

Korean: Twitter
Word Usage Frequency Rank r
Average happiness h_avg
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7
8
9
김정일
한나라당
너무
흑흑
아놔
그걸
같아요
무한
행복한
최고의
최고
씨발
없다고
아침부터
맞다
선물로
크리스마스
사랑해요
유진
무서운
망할
힘들다
왜이리
실시간
대한민국의
축하
사랑합니다
천안함
죄송해요
미친듯이
29일
출연
웃음
자살
욕을
비리
모른다
본인이
친구들이
웃고
축하드립니다
축하해
아픈
사건이
very
좋음
조문
축하합니다
김정일의
무서워서
back
달리
안녕~
^o^
소송
잡아
위해서는
암
변태
짜증
새끼
어렵지
위
운동을
Best
사랑한다
성폭행
일본군
없구나
후후
감사의
행복
커녕
남겨주세요
상상
지키고
선물이
행복해
죽어라
똥



Spanish
Portuguese
English
Indonesian
French
German
Arabic
Russian
Korean
Chinese
h_{avg}
2 3 4 5 6 7 8 9
N=3273 Δ=+0.07 r=0.89
N=3995 Δ=+0.28 r=0.87
N=2206 Δ=+0.34 r=0.82
N=3330 Δ=+0.39 r=0.86
N=2686 Δ=+0.39 r=0.82
N=1306 Δ=+0.51 r=0.83
N=1617 Δ=+0.58 r=0.73
N=801 Δ=+0.55 r=0.79
N=1689 Δ=+0.73 r=0.79
N=3273 Δ=-0.07 r=0.89
N=3592 Δ=+0.20 r=0.87
N=2189 Δ=+0.23 r=0.82
N=2910 Δ=+0.29 r=0.84
N=2547 Δ=+0.31 r=0.81
N=1287 Δ=+0.40 r=0.84
N=1494 Δ=+0.46 r=0.84
N=783 Δ=+0.44 r=0.79
N=1552 Δ=+0.65 r=0.76
N=3995 Δ=-0.28 r=0.87
N=3592 Δ=-0.20 r=0.87
N=2871 Δ=+0.05 r=0.88
N=3526 Δ=+0.12 r=0.86
N=3101 Δ=+0.12 r=0.82
N=1999 Δ=+0.17 r=0.86
N=2011 Δ=+0.21 r=0.87
N=1137 Δ=+0.21 r=0.82
N=2323 Δ=+0.35 r=0.81
N=2206 Δ=-0.34 r=0.82
N=2189 Δ=-0.23 r=0.82
N=2871 Δ=-0.05 r=0.88
N=2130 Δ=+0.04 r=0.79
N=1983 Δ=+0.03 r=0.77
N=1361 Δ=+0.12 r=0.83
N=1246 Δ=+0.12 r=0.85
N=800 Δ=+0.13 r=0.79
N=1404 Δ=+0.32 r=0.77
N=3330 Δ=-0.39 r=0.86
N=2910 Δ=-0.29 r=0.84
N=3526 Δ=-0.12 r=0.86
N=2130 Δ=-0.04 r=0.79
N=2459 Δ=+0.02 r=0.84
N=1275 Δ=+0.09 r=0.77
N=1480 Δ=+0.15 r=0.84
N=772 Δ=+0.12 r=0.79
N=1561 Δ=+0.32 r=0.76
N=2686 Δ=-0.39 r=0.82
N=2547 Δ=-0.31 r=0.81
N=3101 Δ=-0.12 r=0.82
N=1983 Δ=-0.03 r=0.77
N=2459 Δ=-0.02 r=0.84
N=1074 Δ=+0.09 r=0.76
N=1289 Δ=+0.15 r=0.80
N=708 Δ=+0.15 r=0.73
N=1293 Δ=+0.33 r=0.74
N=1306 Δ=-0.51 r=0.83
N=1287 Δ=-0.40 r=0.84
N=1999 Δ=-0.17 r=0.86
N=1361 Δ=-0.12 r=0.83
N=1275 Δ=-0.09 r=0.77
N=1074 Δ=-0.09 r=0.76
N=1300 Δ=+0.03 r=0.83
N=619 Δ=+0.03 r=0.79
N=1321 Δ=+0.23 r=0.80
N=1617 Δ=-0.58 r=0.73
N=1494 Δ=-0.46 r=0.84
N=2011 Δ=-0.21 r=0.87
N=1246 Δ=-0.12 r=0.85
N=1480 Δ=-0.15 r=0.84
N=1289 Δ=-0.15 r=0.80
N=1300 Δ=-0.03 r=0.83
N=679 Δ=+0.04 r=0.80
N=1022 Δ=+0.23 r=0.82
N=801 Δ=-0.55 r=0.79
N=783 Δ=-0.44 r=0.79
N=1137 Δ=-0.21 r=0.82
N=800 Δ=-0.13 r=0.79
N=772 Δ=-0.12 r=0.79
N=708 Δ=-0.15 r=0.73
N=619 Δ=-0.03 r=0.79
N=679 Δ=-0.04 r=0.80
N=934 Δ=+0.18 r=0.81
N=1689 Δ=-0.73 r=0.79
N=1552 Δ=-0.65 r=0.76
N=2323 Δ=-0.35 r=0.81
N=1404 Δ=-0.32 r=0.77
N=1561 Δ=-0.32 r=0.76
N=1293 Δ=-0.33 r=0.74
N=1321 Δ=-0.23 r=0.80
N=1022 Δ=-0.23 r=0.82
N=934 Δ=-0.18 r=0.81

No one understands anything:

A revealing letter and our reply:


The language-dependent relationship between word happiness and frequency

"Language-dependent relationship between word happiness and frequency"
Garcia, Garas, and Schweitzer,
Proc. Natl. Acad. Sci., , , 2015. [7]


Reply to Garcia et al.: Common mistakes in measuring frequency-dependent word characteristics

"Reply to Garcia et al.: Common mistakes in measuring frequency dependent word characteristics"
Dodds et al.,
Proc. Natl. Acad. Sci., , , 2015. [4]

Full version here: http://arxiv.org/abs/1406.3855

Abstract:

The concerns expressed by Garcia *et al.* [7] are misplaced due to a range of misconceptions about word usage frequency, word rank, and expert-constructed word lists such as LIWC [11]. We provide a complete response in our paper's online appendices [3].

LIWC function words are not neutral:

- “greatest” (h_{avg}=7.26),
- “best” (h_{avg}=7.26),
- “unique” (h_{avg}=6.98),
- “negative” (h_{avg}=2.42),
- “worst” (h_{avg}=2.10).

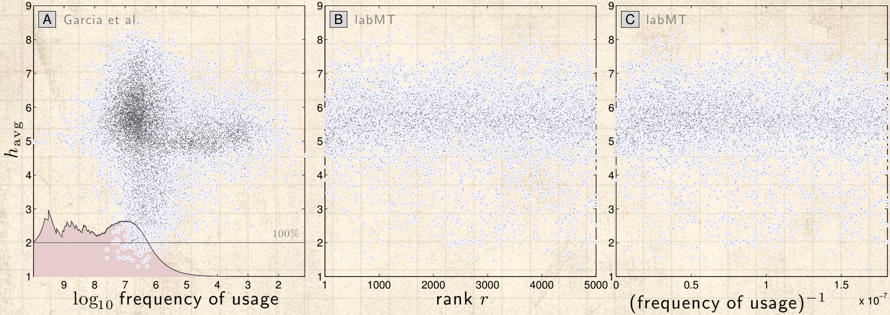
Common scientific sense for text analysis:

Always look at the words.

More LIWC function words:

High	h_{avg}	Neutral	h_{avg}	Low	h_{avg}
billion	7.56	been	5.04	wouldnt	3.86
million	7.38	other	5.04	not	3.86
couple	7.30	into	5.04	shouldn't	3.84
millions	7.26	theyre	5.04	none	3.84
greatest	7.26	it	5.02	haven't	3.82
rest	7.18	some	5.02	wouldn't	3.78
best	7.18	where	5.02	fewer	3.72
equality	7.08	themselves	5.02	lacking	3.71
unique	6.98	im	5.02	won't	3.70
plenty	6.98	quarterly	5.02	wasnt	3.70
truly	6.86	ive	5.02	dont	3.70
hopefully	6.84	because	5.00	don't	3.70
first	6.82	whereas	5.00	down	3.66
plus	6.76	id	5.00	nobody	3.64
well	6.68	til	5.00	doesn't	3.62
greater	6.68	the	4.98	couldnt	3.58
highly	6.60	to	4.98	without	3.54
me	6.58	by	4.98	no	3.48
done	6.54	or	4.98	cant	3.48
extra	6.52	part	4.98	zero	3.44
infinite	6.44	rather	4.98	against	3.40
simply	6.42	its	4.96	never	3.34
equally	6.40	when	4.96	cannot	3.32
sixteen	6.39	perhaps	4.96	lack	3.16
we	6.38	yall	4.96	negative	2.42
soon	6.34	of	4.94	worst	2.10

The jellyfish knows:



Scatterplot of h_{avg} as a function of word usage frequency for the English Google Books word list generated by Garcia *et al.*. Uncontrolled subsampling of lower frequency words yields a lexicon that is not statistically representative of any natural language corpus. The lower curve provides a coarse estimate of cumulative lexicon coverage as a function of usage frequency f using Zipf's law $f_r \sim f_1 r^{-1}$ inverted as $r \sim f_1/f_r$. The rapid drop off begins at around rank 5000, the involved lexicon size for Google Books in labMT [3, 6]. **B.** and Scatterplot of h_{avg} as a function of rank r for the 5000 words for Google Books contributing to labMT, the basis of our jellyfish plots [3]. **C.** Same data as **B** plotted against f. Linear regression fits for the first two scatterplots are $h_{\text{avg}} \simeq 0.089 \log_{10} f + 4.85$ and $h_{\text{avg}} \simeq -3.04 \times 10^{-5} r + 5.62$ (as reported in [3]). Note difference in signs, and the far weaker trend for the statistically appropriate regression against rank in **B**. Pearson correlation coefficients: +0.105, -0.042, and -0.043 with p-values 6.15$\times 10^{-26}$, 3.03$\times 10^{-3}$, and 2.57$\times 10^{-3}$. Spearman correlation coefficients: +0.201, -0.013, and -0.013 with p-values 6.37$\times 10^{-92}$, 0.350, and 0.350.

Nutshell:

- Linguistic positivity bias holds for 10 major languages.
- Spread across 24 corpora: books, news, social media, movie titles, ...
- Languages and evaluating groups spread around the world.
- Diverse in language origins.
- Language appears to reflect social, cooperative tendency of people.
- Negative emotion is more variable—must be specific, Tolstoyfully.

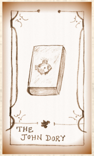
Corpus:	# Words	Reference(s)
English: Twitter	5000	[?, 6]
English: Google Books Project	5000	[10]
English: The New York Times	5000	[12]
English: Music lyrics	5000	[5]
Portuguese: Google Web Crawl	7133	[?]
Portuguese: Twitter	7119	[?]
Spanish: Google Web Crawl	7189	[?]
Spanish: Twitter	6415	[?]
Spanish: Google Books Project	6379	[10]
French: Google Web Crawl	7056	[?]
French: Twitter	6569	[?]
French: Google Books Project	6192	[10]
Arabic: Movie and TV subtitles	9999	MITRE
Indonesian: Twitter	7044	[?]
Indonesian: Movie subtitles	6726	MITRE
Russian: Twitter	6575	[?]
Russian: Google Books Project	5980	[10]
Russian: Movie and TV subtitles	6186	[?]
German: Google Web Crawl	6902	[?]
German: Twitter	6459	[?]
German: Google Books Project	6097	[10]
Korean: Twitter	6728	[?]
Korean: Movie subtitles	5389	MITRE
Chinese: Google Books Project	10000	[10]

Language	Participants' location(s)	# of participants	Average words scored
English	US, India	384	1302
German	Germany	196	2551
Indonesian	Indonesia	146	3425
Russian	Russia	125	4000
Arabic	Egypt	185	2703
French	France	179	2793
Spanish	Mexico	236	2119
Portuguese	Brazil	208	2404
Simplified Chinese	China	128	3906
Korean	Korea, US	109	4587

Number and main country/countries of location for participants evaluating the 10,000 common words for each of the 10 languages we studied. Also recorded is the average number of words evaluated by each participant (rounded to the nearest integer). We note that each word received 50 evaluations from distinct individuals. The English word list was evaluated via Mechanical Turk for our initial study [9]. The nine languages evaluated through Appen-Butler Hill yielded a higher participation rate likely due to better pay and the organization's quality of service.

We used the services of Appen Butler Hill (http://www.appen.com) for all word evaluations excluding English, for which we had earlier employed Mechanical Turk (https://www.mturk.com/ [9]).

English instructions were translated to all other languages and given to participants along with survey questions, and an example of the English instruction page is below. Non-english language experiments were conducted through a custom interactive website built by Appen Butler Hill, and all participants were required to pass a stringent aural proficiency test in their own language.



Measuring the Happiness of Words

Our overall aim is to assess how people feel about individual words. With this particular survey, we are focusing on the dual emotions of sadness and happiness. You are to rate 100 individual words on a 9 point unhappy-happy scale.

Please consider each word carefully. If we determine that your ratings are randomly or otherwise inappropriately selected, or that any questions are left unanswered, we may not approve your work. These words were chosen based on their common usage. As a result, a small portion of words may be offensive to some people, written in a different language, or nonsensical.

Before completing the word ratings, we ask that you answer a few short demographic questions. We expect the entire survey to require 10 minutes of your time. Thank you for participating!

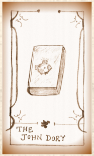
Example:

sunshine

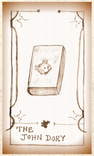
Read the word and click on the face that best corresponds to your emotional response.

Demographic Questions

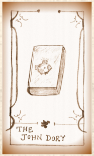
1. What is your gender? (Male/Female)
2. What is your age? (Free text)
3. Which of the following best describes your highest achieved education level? Some High School, High School Graduate, Some college, no degree, Associates degree, Bachelors degree, Graduate degree (Masters, Doctorate, etc.)
4. What is the total income of your household?
5. Where are you from originally?
6. Where do you live currently?
7. Is _______ your first language? (Yes/No) If it is not, please specify what your first language is.
8. Do you have any comments or suggestions? (Free text)



Of our 24 corpora, we received 17 already parsed by the source: the Google Books Project (6 corpora), the Google Web Crawl (8 corpora), and Movie and TV subtitles (3 corpora). For the other 7 corpora (Twitter, New York Times, and Music Lyrics), we extracted words by standard white space separation (more on Twitter below). We acknowledge the many complications with inflections and variable orthography. We have found merit in not collapsing related words, which would require a more sophisticated treatment going beyond the present paper's bounds. Moreover, we have observed that allowing, say, different conjugation of verbs to stand in our corpora is valuable as human evaluations of such have proved to be distinguishable (e.g., present versus past tense [6]).



Twitter was easily the most variable and unruly of our text sources and required additional treatment. We first checked if a string contains at least one valid utf8 letter, discarding if not. Next we filtered out strings containing invisible control characters, as these symbols can be problematic. We ignored all strings that start with $<$ and end with $>$ (generally html code). We ignored strings with a leading @ or &, or either preceded with standard punctuation (e.g., Twitter ID's), but kept hashtags. We also removed all strings starting with www. or http: or end in .com (all websites). We stripped the remaining strings of standard punctuation, and we replaced all double quotes (") by single quotes ('). Finally, we converted all Latin alphabet letters to lowercase.



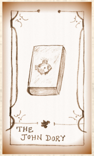
Tokenization example:

Term	count
love	10
LoVE	5
love!	2
#love	3
.love	2
@love	1
love87	1

$\rightarrow$

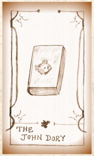
Term	count
love	19
#love	3
love87	1

The term ‘@love’ is discarded, and all other terms map to either ‘love’ or ‘love87’.



There is no single, principled way to merge corpora to create an ordered list of words for a given language. For example, it is impossible to weight the most commonly used words in the New York Times against those of Twitter. Nevertheless, we are obliged to choose some method for doing so to facilitate comparisons across languages and for the purposes of building adaptable linguistic instruments.

For each language where we had more than one corpus, we created a single quasi-ranked word list by finding the smallest integer r such that the union of all words with rank $\leq r$ in at least one corpus formed a set of at least 10,000 words.



	Spanish	Portuguese	English	Indonesian	French	German	Arabic
Spanish	1.00, 0.00	1.01, 0.03	1.06, -0.07	1.22, -0.88	1.11, -0.24	1.22, -0.84	1.13, -0.22
Portuguese	0.99, -0.03	1.00, 0.00	1.04, -0.03	1.22, -0.97	1.11, -0.33	1.21, -0.86	1.09, -0.08
English	0.94, 0.06	0.96, 0.03	1.00, 0.00	1.13, -0.66	1.06, -0.23	1.16, -0.75	1.05, -0.10
Indonesian	0.82, 0.72	0.82, 0.80	0.88, 0.58	1.00, 0.00	0.92, 0.48	0.99, 0.06	0.89, 0.71
French	0.90, 0.22	0.90, 0.30	0.94, 0.22	1.09, -0.52	1.00, 0.00	1.08, -0.44	0.99, 0.12
German	0.82, 0.69	0.83, 0.71	0.86, 0.65	1.01, -0.06	0.92, 0.41	1.00, 0.00	0.91, 0.61
Arabic	0.88, 0.19	0.92, 0.08	0.95, 0.10	1.12, -0.80	1.01, -0.12	1.10, -0.68	1.00, 0.00
Russian	0.76, 0.88	0.80, 0.75	0.83, 0.75	0.98, -0.04	0.89, 0.45	0.93, 0.24	0.89, 0.56
Korean	0.62, 1.70	0.62, 1.81	0.66, 1.67	0.77, 1.17	0.73, 1.37	0.78, 1.12	0.71, 1.53
Chinese	0.63, 1.46	0.63, 1.51	0.68, 1.43	0.75, 1.07	0.71, 1.26	0.76, 1.03	0.70, 1.41

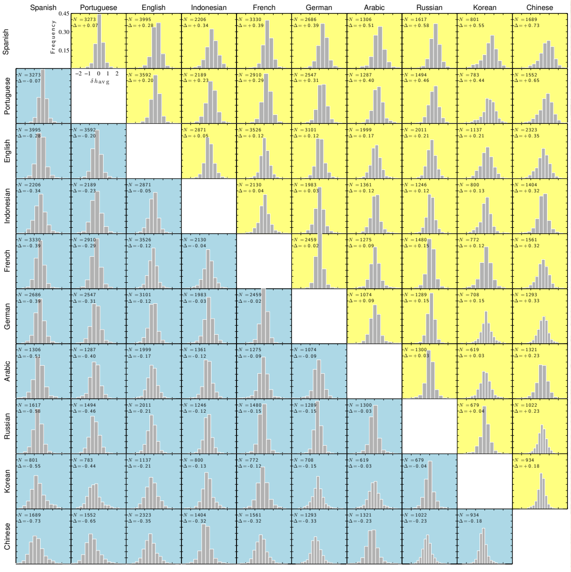
Reduced Major Axis (RMA) regression fits for row language as a linear function of the column language: $h_{\text{avg}}^{\text{(row)}}(w) = m\,h_{\text{avg}}^{\text{(column)}}(w) + c$ where w indicates a translation-stable word. Each entry in the table contains the coefficient pair m and c. We use RMA regression, also known as Standardized Major Axis linear regression, because of its accommodation of errors in both variables.

	Spanish	Portuguese	English	Indonesian	French	German	Arabic	Russian
Spanish	1.00	0.89	0.87	0.82	0.86	0.82	0.83	0.73
Portuguese	0.89	1.00	0.87	0.82	0.84	0.81	0.84	0.84
English	0.87	0.87	1.00	0.88	0.86	0.82	0.86	0.87
Indonesian	0.82	0.82	0.88	1.00	0.79	0.77	0.83	0.85
French	0.86	0.84	0.86	0.79	1.00	0.84	0.77	0.84
German	0.82	0.81	0.82	0.77	0.84	1.00	0.76	0.80
Arabic	0.83	0.84	0.86	0.83	0.77	0.76	1.00	0.83
Russian	0.73	0.84	0.87	0.85	0.84	0.80	0.83	1.00
Korean	0.79	0.79	0.82	0.79	0.79	0.73	0.79	0.80
Chinese	0.79	0.76	0.81	0.77	0.76	0.74	0.80	0.82

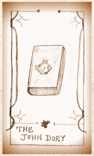
Pearson correlation coefficients for translation-stable words for all language pairs. All p-values are $< 10^{-118}$.

	Spanish	Portuguese	English	Indonesian	French	German	Arabic	Russian
Spanish	1.00	0.85	0.83	0.77	0.81	0.77	0.75	0.74
Portuguese	0.85	1.00	0.83	0.77	0.78	0.77	0.77	0.81
English	0.83	0.83	1.00	0.82	0.80	0.78	0.78	0.81
Indonesian	0.77	0.77	0.82	1.00	0.72	0.72	0.76	0.77
French	0.81	0.78	0.80	0.72	1.00	0.80	0.67	0.79
German	0.77	0.77	0.78	0.72	0.80	1.00	0.69	0.76
Arabic	0.75	0.77	0.78	0.76	0.67	0.69	1.00	0.74
Russian	0.74	0.81	0.81	0.77	0.79	0.76	0.74	1.00
Korean	0.74	0.75	0.75	0.71	0.71	0.64	0.69	0.70
Chinese	0.68	0.66	0.70	0.71	0.64	0.62	0.68	0.66

Spearman correlation coefficients for translation-stable words. All p-values are $< 10^{-82}$.



Histograms of the change in average happiness for translation-stable words between each language pair. The largest deviations correspond to strong changes in a word's perceived primary meaning (e.g., 'lying' and 'acostado'). The inset quantities are N, the number of translation-stable words, and Δ is the average difference in translation-stable word happiness between the row language and column language.

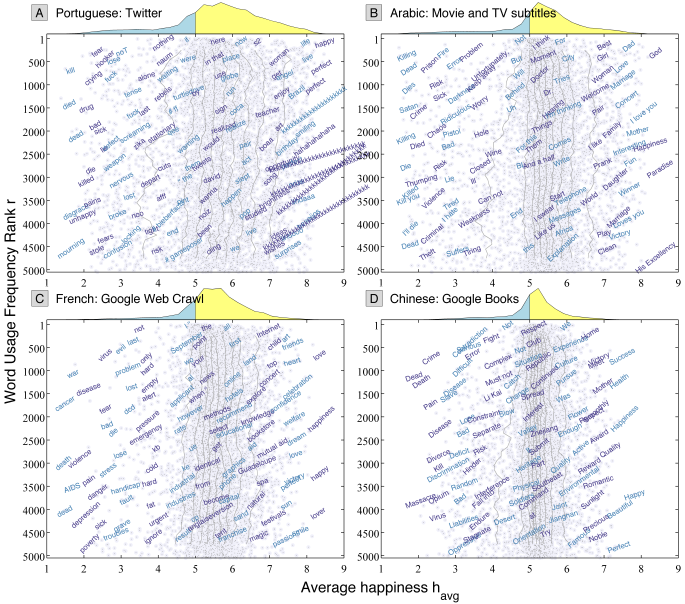


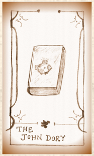
Language: Corpus	ρ_p	p-value	ρ_s	p-value	α	β
Spanish: Google Web Crawl	-0.114	3.38×10^{-22}	-0.090	1.85×10^{-14}	-5.55×10^{-5}	6.10
Spanish: Google Books	-0.040	1.51×10^{-3}	-0.016	1.90×10^{-1}	-2.28×10^{-5}	5.90
Spanish: Twitter	-0.048	1.14×10^{-4}	-0.032	1.10×10^{-2}	-3.10×10^{-5}	5.94
Portuguese: Google Web Crawl	-0.085	6.33×10^{-13}	-0.060	3.23×10^{-7}	-3.98×10^{-5}	5.96
Portuguese: Twitter	-0.041	5.98×10^{-4}	-0.030	1.15×10^{-2}	-2.40×10^{-5}	5.73
English: Google Books	-0.042	3.03×10^{-3}	-0.013	3.50×10^{-1}	-3.04×10^{-5}	5.62
English: New York Times	-0.056	6.93×10^{-5}	-0.044	1.99×10^{-3}	-4.17×10^{-5}	5.61
German: Google Web Crawl	-0.096	1.11×10^{-15}	-0.082	6.75×10^{-12}	-3.67×10^{-5}	5.65
French: Google Web Crawl	-0.105	9.20×10^{-19}	-0.080	1.99×10^{-11}	-4.50×10^{-5}	5.68
English: Twitter	-0.097	6.56×10^{-12}	-0.103	2.37×10^{-13}	-7.78×10^{-5}	5.67
Indonesian: Movie subtitles	-0.039	1.48×10^{-3}	-0.063	2.45×10^{-7}	-2.04×10^{-5}	5.45
German: Twitter	-0.054	1.47×10^{-5}	-0.036	4.02×10^{-3}	-2.51×10^{-5}	5.58
Russian: Twitter	-0.052	2.38×10^{-5}	-0.028	2.42×10^{-2}	-2.55×10^{-5}	5.52
French: Google Books	-0.043	6.80×10^{-4}	-0.030	1.71×10^{-2}	-2.31×10^{-5}	5.49
German: Google Books	-0.003	8.12×10^{-1}	+0.014	2.74×10^{-1}	-1.38×10^{-6}	5.45
French: Twitter	-0.049	6.08×10^{-5}	-0.023	6.31×10^{-2}	-2.54×10^{-5}	5.54
Russian: Movie and TV subtitles	-0.029	2.36×10^{-2}	-0.033	9.17×10^{-3}	-1.57×10^{-5}	5.43
Arabic: Movie and TV subtitles	-0.045	7.10×10^{-6}	-0.029	4.19×10^{-3}	-1.66×10^{-5}	5.44
Indonesian: Twitter	-0.051	2.14×10^{-5}	-0.018	1.24×10^{-1}	-2.50×10^{-5}	5.46
Korean: Twitter	-0.032	8.29×10^{-3}	-0.016	1.91×10^{-1}	-1.24×10^{-5}	5.38
Russian: Google Books	+0.030	2.09×10^{-2}	+0.070	5.08×10^{-8}	$+1.20\times10^{-5}$	5.35
English: Music Lyrics	-0.073	2.53×10^{-7}	-0.081	1.05×10^{-8}	-6.12×10^{-5}	5.45
Korean: Movie subtitles	-0.187	8.22×10^{-44}	-0.180	2.01×10^{-40}	-9.66×10^{-5}	5.41
Chinese: Google Books	-0.067	1.48×10^{-11}	-0.050	5.01×10^{-7}	-1.72×10^{-5}	5.21

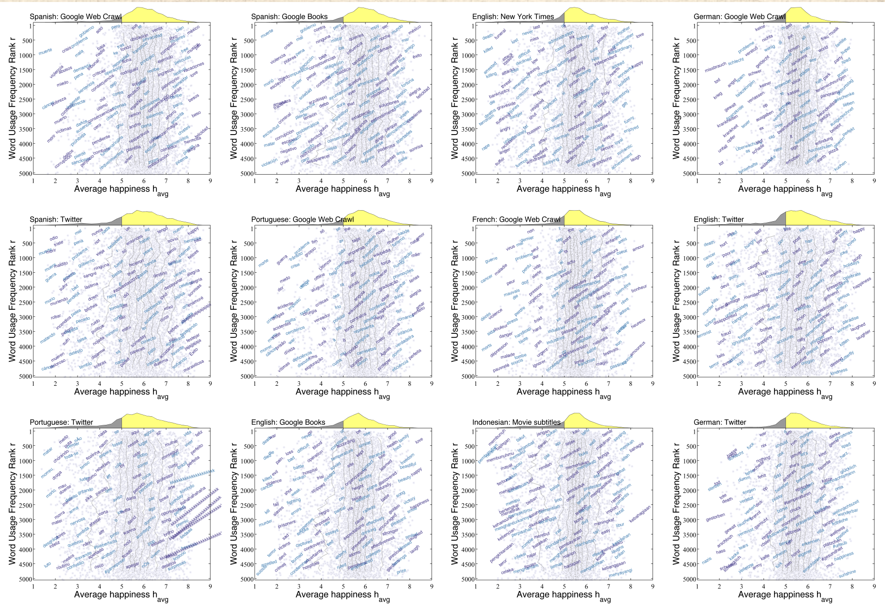
Pearson correlation coefficients and p-values, Spearman correlation coefficients and p-values, and linear fit coefficients, for average word happiness h_{avg} as a function of word usage frequency rank r. We use the fit is $h_{\text{avg}} = \alpha r + \beta$ for the most common 5000 words in each corpora, determining α and β via ordinary least squares, and order languages by the median of their average word happiness scores (descending). We note that stemming of words may affect these estimates.

Language: Corpus	ρ_{p}	p-value	ρ_{s}	p-value	α	β
Portuguese: Twitter	+0.090	2.55×10^{-14}	+0.095	1.28×10^{-15}	1.19×10^{-5}	1.29
Spanish: Twitter	+0.097	8.45×10^{-15}	+0.104	5.92×10^{-17}	1.47×10^{-5}	1.26
English: Music Lyrics	+0.129	4.87×10^{-20}	+0.134	1.63×10^{-21}	2.76×10^{-5}	1.33
English: Twitter	+0.007	6.26×10^{-1}	+0.012	4.11×10^{-1}	1.47×10^{-6}	1.35
English: New York Times	+0.050	4.56×10^{-4}	+0.044	1.91×10^{-3}	9.34×10^{-6}	1.32
Arabic: Movie and TV subtitles	+0.101	7.13×10^{-24}	+0.101	3.41×10^{-24}	9.41×10^{-6}	1.01
English: Google Books	+0.180	1.68×10^{-37}	+0.176	4.96×10^{-36}	3.36×10^{-5}	1.27
Spanish: Google Books	+0.066	1.23×10^{-7}	+0.062	6.53×10^{-7}	9.17×10^{-6}	1.26
Indonesian: Movie subtitles	+0.026	3.43×10^{-2}	+0.027	2.81×10^{-2}	2.87×10^{-6}	1.12
Russian: Movie and TV subtitles	+0.083	7.60×10^{-11}	+0.075	3.28×10^{-9}	1.06×10^{-5}	0.89
French: Twitter	+0.072	4.77×10^{-9}	+0.076	8.94×10^{-10}	1.07×10^{-5}	1.05
Indonesian: Twitter	+0.072	1.17×10^{-9}	+0.072	1.73×10^{-9}	8.16×10^{-6}	1.12
French: Google Books	+0.090	1.02×10^{-12}	+0.085	1.67×10^{-11}	1.25×10^{-5}	1.02
Russian: Twitter	+0.055	6.83×10^{-6}	+0.053	1.67×10^{-5}	7.39×10^{-6}	0.91
Spanish: Google Web Crawl	+0.119	4.45×10^{-24}	+0.106	2.60×10^{-19}	1.45×10^{-5}	1.23
Portuguese: Google Web Crawl	+0.093	4.06×10^{-15}	+0.083	2.91×10^{-12}	1.07×10^{-5}	1.26
German: Twitter	+0.051	4.45×10^{-5}	+0.050	5.15×10^{-5}	7.39×10^{-6}	1.15
French: Google Web Crawl	+0.104	2.12×10^{-18}	+0.088	9.64×10^{-14}	1.27×10^{-5}	1.01
Korean: Movie subtitles	+0.171	1.39×10^{-36}	+0.185	8.85×10^{-43}	2.58×10^{-5}	0.88
German: Google Books	+0.157	6.06×10^{-35}	+0.162	4.96×10^{-37}	2.17×10^{-5}	1.03
Korean: Twitter	+0.056	4.07×10^{-6}	+0.062	4.25×10^{-7}	6.98×10^{-6}	0.93
German: Google Web Crawl	+0.099	2.05×10^{-16}	+0.085	1.18×10^{-12}	1.20×10^{-5}	1.07
Chinese: Google Books	+0.099	3.07×10^{-23}	+0.097	3.81×10^{-22}	8.70×10^{-6}	1.16
Russian: Google Books	+0.187	5.15×10^{-48}	+0.177	2.24×10^{-43}	2.28×10^{-5}	0.81

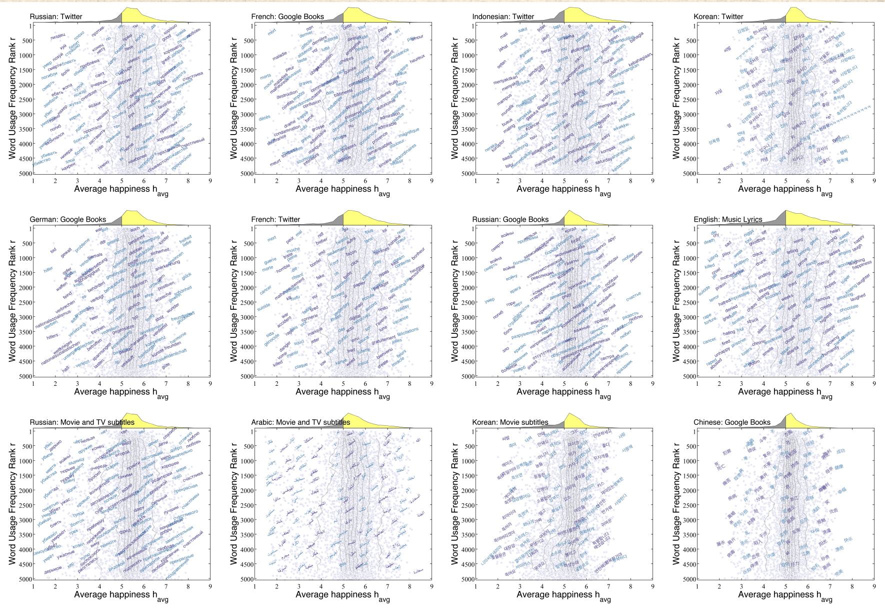
Pearson correlation coefficients and p-values, Spearman correlation coefficients and p-values, and linear fit coefficients for standard deviation of word happiness h_{std} as a function of word usage frequency rank r. We consider the fit is $h_{\text{std}} = \alpha r + \beta$ for the most common 5000 words in each corpora, determining α and β via ordinary least squares, and order corpora according to their emotional variance (descending).


A Portuguese: Twitter
B Arabic: Movie and TV subtitles
C French: Google Web Crawl
D Chinese: Google Books
Word Usage Frequency Rank r
Average happiness h_avg

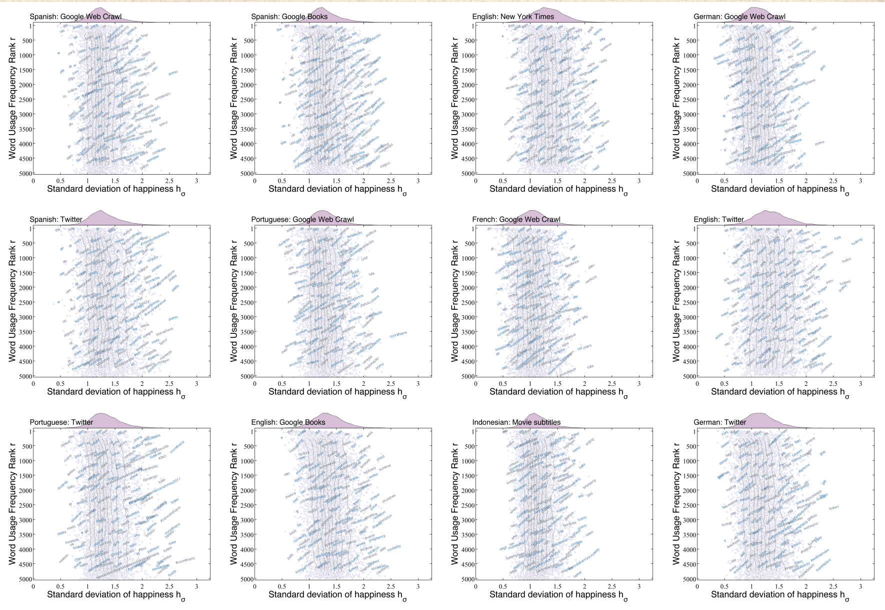

THE
JOHN DORY

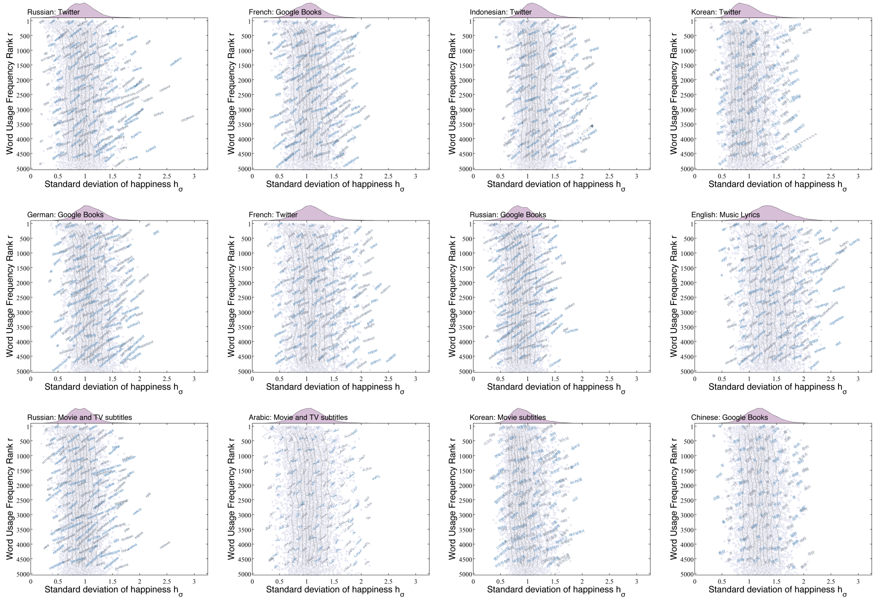


Spanish: Google Web Crawl
Spanish: Google Books
English: New York Times
German: Google Web Crawl
Spanish: Twitter
Portuguese: Google Web Crawl
French: Google Web Crawl
English: Twitter
Portuguese: Twitter
English: Google Books
Indonesian: Movie subtitles
German: Twitter
Word Usage Frequency Rank r
Average happiness h_{avg}

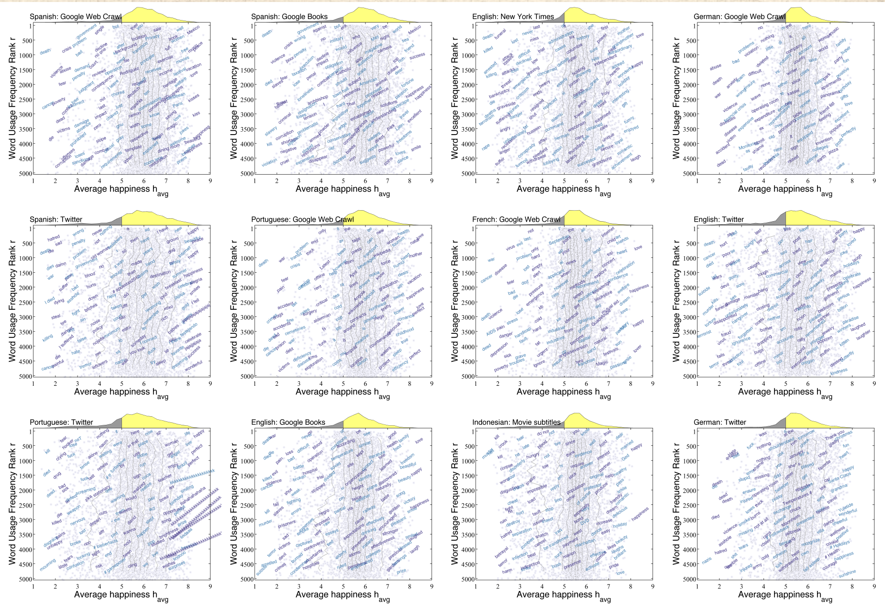


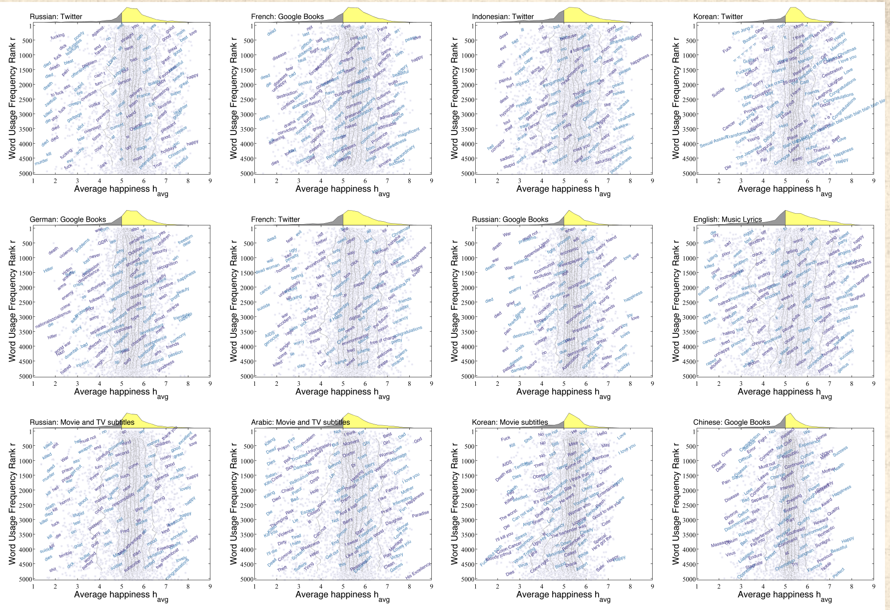
Russian: Twitter
French: Google Books
Indonesian: Twitter
Korean: Twitter
German: Google Books
French: Twitter
Russian: Google Books
English: Music Lyrics
Russian: Movie and TV subtitles
Arabic: Movie and TV subtitles
Korean: Movie subtitles
Chinese: Google Books
Word Usage Frequency Rank r
Average happiness h_{avg}


Spanish: Google Web Crawl
Spanish: Google Books
English: New York Times
German: Google Web Crawl
Spanish: Twitter
Portuguese: Google Web Crawl
French: Google Web Crawl
English: Twitter
Portuguese: Twitter
English: Google Books
Indonesian: Movie subtitles
German: Twitter
Word Usage Frequency Rank r
Standard deviation of happiness h_σ

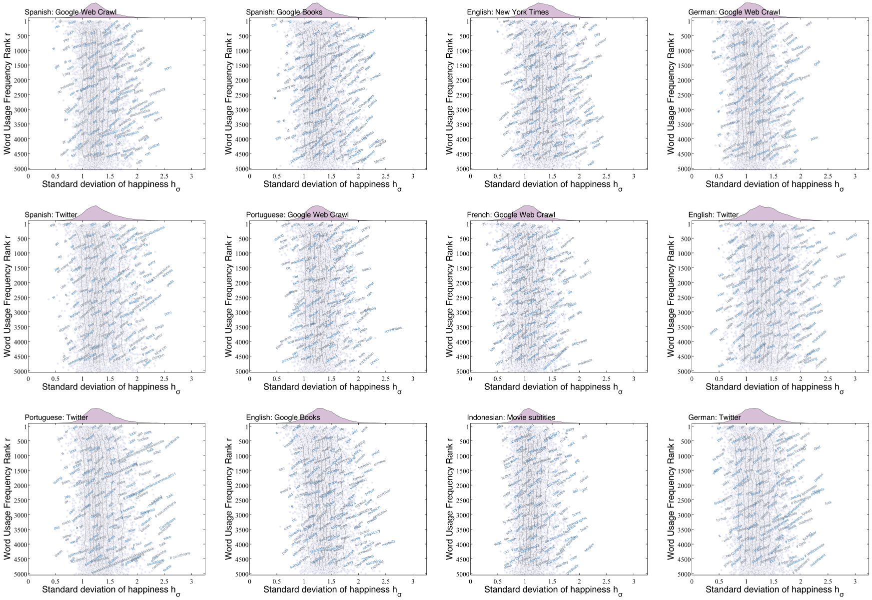


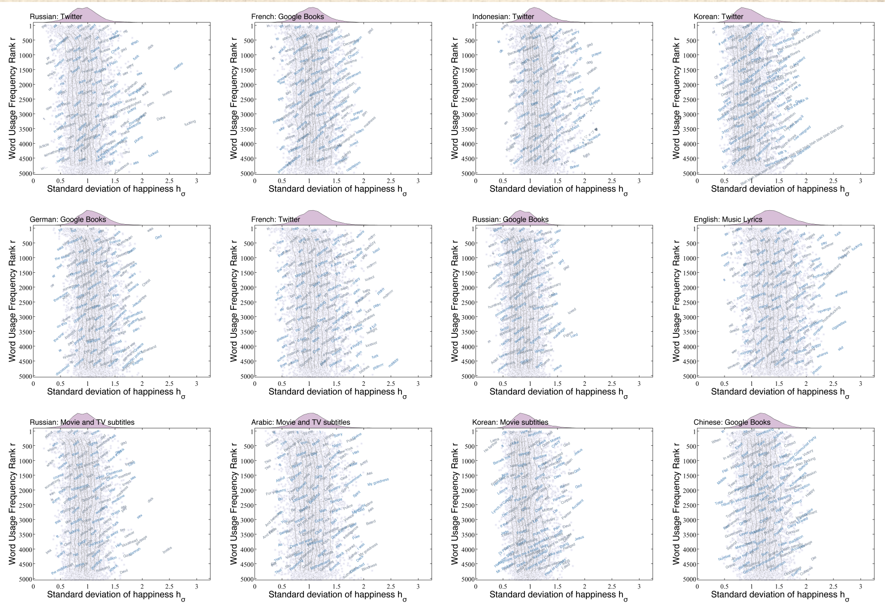
Russian: Twitter
French: Google Books
Indonesian: Twitter
Korean: Twitter
German: Google Books
French: Twitter
Russian: Google Books
English: Music Lyrics
Russian: Movie and TV subtitles
Arabic: Movie and TV subtitles
Korean: Movie subtitles
Chinese: Google Books
Word Usage Frequency Rank r
Standard deviation of happiness h_σ


Spanish: Google Web Crawl
Spanish: Google Books
English: New York Times
German: Google Web Crawl
Spanish: Twitter
Portuguese: Google Web Crawl
French: Google Web Crawl
English: Twitter
Portuguese: Twitter
English: Google Books
Indonesian: Movie subtitles
German: Twitter
Word Usage Frequency Rank r
Average happiness h avg



Russian: Twitter
French: Google Books
Indonesian: Twitter
Korean: Twitter
German: Google Books
French: Twitter
Russian: Google Books
English: Music Lyrics
Russian: Movie and TV subtitles
Arabic: Movie and TV subtitles
Korean: Movie subtitles
Chinese: Google Books
Word Usage Frequency Rank r
Average happiness h_{avg}


Spanish: Google Web Crawl
Spanish: Google Books
English: New York Times
German: Google Web Crawl
Spanish: Twitter
Portuguese: Google Web Crawl
French: Google Web Crawl
English: Twitter
Portuguese: Twitter
English: Google Books
Indonesian: Movie subtitles
German: Twitter
Word Usage Frequency Rank r
Standard deviation of happiness h_σ

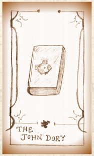


Russian: Twitter
French: Google Books
Indonesian: Twitter
Korean: Twitter
German: Google Books
French: Twitter
Russian: Google Books
English: Music Lyrics
Russian: Movie and TV subtitles
Arabic: Movie and TV subtitles
Korean: Movie subtitles
Chinese: Google Books
Word Usage Frequency Rank r
Standard deviation of happiness h_σ

References I

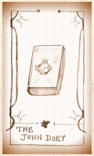
[1] J. M. Diamond.
Guns, Germs, and Steel.
W. W. Norton & Company, 1997.

[2] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth.
Human language reveals a universal positivity bias.
Proc. Natl. Acad. Sci., 112(8):2389–2394, 2015.
Available online at http://www.pnas.org/content/112/8/2389. pdf



References II

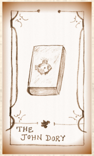
[3] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth.
Human language reveals a universal positivity bias.
Proc. Natl. Acad. Sci., 112(8):2389–2394, 2015.
Available online at http://www.pnas.org/content/112/8/2389; online appendices: http://compstorylab.org/share/papers/dodds2014a/.



References III

[4] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth.
Reply to garcia et al.: Common mistakes in measuring frequency dependent word characteristics.
Proc. Natl. Acad. Sci., 2015.
Available online at http://www.pnas.org/content/early/2015/05/20/1505647112. pdf

[5] P. S. Dodds and C. M. Danforth.
Measuring the happiness of large-scale written expression: songs, blogs, and presidents.
Journal of Happiness Studies, 2009.
doi:10.1007/s10902-009-9150-9. pdf

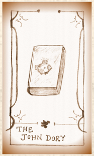


References IV

[6] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth.
Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter.
PLoS ONE, 6:e26752, 2011. pdf

[7] D. Garcia, A. Garas, and F. Schweitzer.
Language-dependent relationship between word happiness and frequency.
Proc. Natl. Acad. Sci., 2015.
doi: 10.1073/pnas.1502909112. pdf

[8] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds.
Positivity of the English language.
PLoS ONE, 7:e29484, 2012. pdf

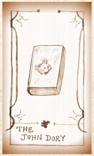


References V

[9] I. M. Kloumann, C. M. Danforth, K. D. Harris, C. A. Bliss, and P. S. Dodds.
Positivity of the English language.
PLoS ONE, 7:e29484, 2012. pdf

[10] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman.
Quantitative analysis of culture using millions of digitized books.
Science Magazine, 331:176–182, 2011. pdf

[11] J. W. Pennebaker, R. J. Booth, and M. E. Francis.
Linguistic Inquiry and Word Count: LIWC 2007.
at http://bit.ly/S1Dk2L, accessed May 15, 2014., 2007.



References VI

[12] E. Sandhaus.
The New York Times Annotated Corpus.
Linguistic Data Consortium, Philadelphia, 2008.
Available online at:
https://doi.org/10.35111/77ba-9x74.