

The Expression of Emotions in 20th Century Books

Alberto Acerbi^{1,2*}, Vasileios Lampos³, Philip Garnett⁴, R. Alexander Bentley¹

1 Department of Archaeology and Anthropology, University of Bristol, Bristol, United Kingdom, **2** Centre for the Study of Cultural Evolution, Stockholm University, Stockholm, Sweden, **3** Department of Computer Science, University of Sheffield, Sheffield, United Kingdom, **4** Department of Anthropology, Durham University, Durham, United Kingdom

Abstract

We report here trends in the usage of “mood” words, that is, words carrying emotional content, in 20th century English language books, using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative moods, underlain by a general decrease in the use of emotion-related words through time. Finally, we show that, in books, American English has become decidedly more “emotional” than British English in the last half-century, as a part of a more general increase of the stylistic divergence between the two variants of English language.

Citation: Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The Expression of Emotions in 20th Century Books. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

Editor: Sune Lehmann, Technical University of Denmark, Denmark

Received: July 26, 2012; **Accepted:** February 11, 2013; **Published:** March 20, 2013

Copyright: © 2013 Acerbi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was partially supported by the Leverhulme Trust “Tipping Points” program. A. Acerbi was initially supported by the “Uniquely Human” project funded by the Swedish Research Council, and then by a Newton International Fellowship. V. Lampos acknowledges the support from the TrendMiner project (EU-FP7-ICT n.287863). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Co-author R. Alexander Bentley is a PLOS ONE Editorial Board member. This does not alter the authors’ adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: alberto.acerbi@gmail.com

Introduction

Studies of cultural change and evolution have been transformed recently by a new level of accessibility and volume of electronic data concerning human behavior [1]. Among these, new studies of word usage can assess mass everyday public preferences that may be missed in mainstream history and politics [2–8]. For example, interest in mass public ‘mood’ was sparked by claims that Twitter mood words can predict stock market trends [9], and, in general, sentiment analysis of Twitter stream is a growing field of research [10–12]. This has been paired by findings of surprisingly straightforward patterns about word usage online, such as S-curves of adoption [13–15], or traders whose instant messaging patterns correlate with their financial success [16–18].

While these studies focus on the recent, short-time scale of online media, one of the exciting challenges of the ‘Big Data’ agenda is to address cultural dynamics at longer time scales. Language itself is a remarkably long-lived phenomenon, with most of the common words in a language having been passed down through multiple generations for centuries or millennia [3,4,19]. While studies of word usage in different online communities [13] can achieve unprecedented sample sizes, long-term temporal dynamics and diversity are exhibited by the history and inferred prehistory of human languages [20–23].

In a novel approach, Hughes et al. [7] investigated past centuries of books (using the Project Gutenberg Digital Library) through the frequencies of ‘content-free’ words, that is words that carry little or not meaning on their own, but acquire it according to the context in which they are used, such as “to be” verbs, conjunctions (e.g. *and*, *but*), articles (e.g. *the*), pronouns (e.g. *you*, *us*) and prepositions (e.g. *about*, *within*). Representing authors by a vector of normalised content-free word frequencies used in their works, Hughes et al. [7] found temporal regimes of similarity

among authors (1784–1829, 1825–1870, 1866–1911, and 1907–1952) that provided a fascinating comparison with well-known historical genres of literature. In evolutionary terms, words with specific content are selected to serve a more direct purpose, but the content-free words subject to random drift mark differences in stylistic genre.

If content-free words are a good proxy for stylistic change, we may add in increments of ‘content’ and see how patterns change. For example, using Google’s Ngram database [2], Twenge et al. [8] found that the summed frequencies of ‘individualistic’ words (e.g., *independent*, *individual*, *unique*, *self*, *solitary*, *personal*) significantly increased in American books between 1960 and 2008, while ‘communal’ words (e.g., *communal*, *team*, *collective*, *village*, *group*, *union*) did not. During those same 48 years, individualistic phrases (e.g., *all about me*, *I get what I want*) also increased in frequency compared to communal phrases (e.g., *band together*, *united we stand*) [8].

Here we analyze trends in the past century of mood words in books, using Google’s Ngram database. Google’s Ngram database represents a 4% digitally-scanned sample of several centuries of books, for a total of 5,195,769 volumes [2]. The corpus contains texts in different languages, and, for English, a further distinction is made between American English and British English (according to the country of publication, i.e. United States versus Great Britain). Additionally, a subset of English texts collects only fiction books. Titles of books present in the corpus are not available because of copyright reasons [2]. The corpus gives information on how many times, in a given year, an 1-gram or an n-gram is used, where an 1-gram is a string of characters uninterrupted by space (i.e. a word, but also numbers, typos, etc.) and an n-gram is a sequence of *n* 1-grams.

We make use of six unique lists of terms (see Methods) to characterize mood categories labeled as Anger, Disgust, Fear, Joy, Sadness, and Surprise. These mood word lists have previously

been applied on a study of U.K. Twitter content, which showed that changes in these mood word frequencies identified real-world events such as the unexpected deaths of popular personas, public unrest, or natural disasters [24]. We extend the time scale of this analysis by tracking mood word frequencies through the past century of Google book data. We find a general decrease in the use of mood terms through time, which underlies a distinct increase in emotional word usage in American books versus British books in the last half century.

Results

Our analysis yielded three main results. First, we can distinguish between ‘happy’ and ‘sad’ periods in the data, plotting the differences between *z*-scores (see Methods) for Joy and Sadness in the 1-grams English data set. Figure 1 shows that moods tracked broad historical trends, including a ‘sad’ peak corresponding to Second World War, and two ‘happy’ peaks, one in the 1920’s and the other in the 1960’s. In more recent years we can see a ‘sad’ period starting from the 1970’s, with an increase in ‘happiness’ in the last years of the data set. Interestingly, the First World War does not seem to register a particular change in mood words (Figure 1).

Our second finding is a clear decrease in the overall use of mood words through time (Figure 2). We performed checks to confirm that the overall decrease in mood word frequency in the data is not merely a reflection of, for example, greater numbers of technically-oriented or scientific books through time. Although the Ngram database does not give an explicit breakdown of book subject categories [2], we analyzed the same mood word lists on Google’s 1-grams English Fiction data set, which contains only works of fiction and literary criticism. In support of a real decrease in

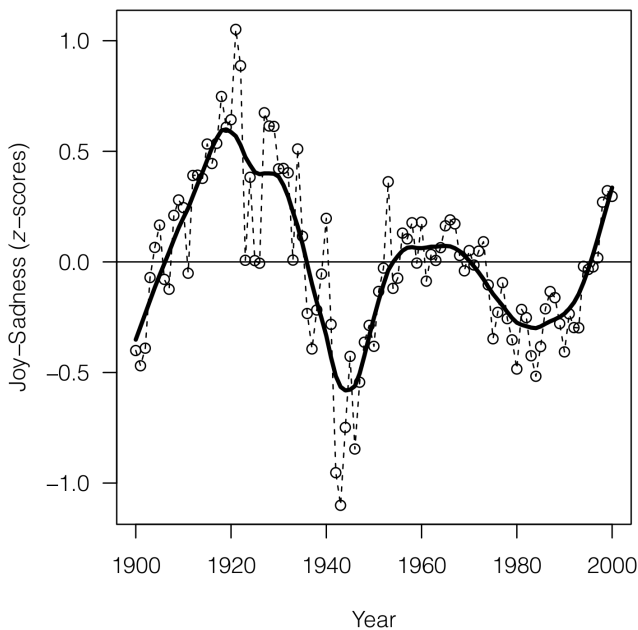


Figure 1. Historical periods of positive and negative moods. Difference between *z*-scores of Joy and Sadness for years from 1900 to 2000 (raw data and smoothed trend). Values above zero indicate generally ‘happy’ periods, and values below the zero indicate generally ‘sad’ periods. Values are smoothed using Friedman’s ‘super smoother’ through R function `supsmu()` [47]. doi:10.1371/journal.pone.0059030.g001

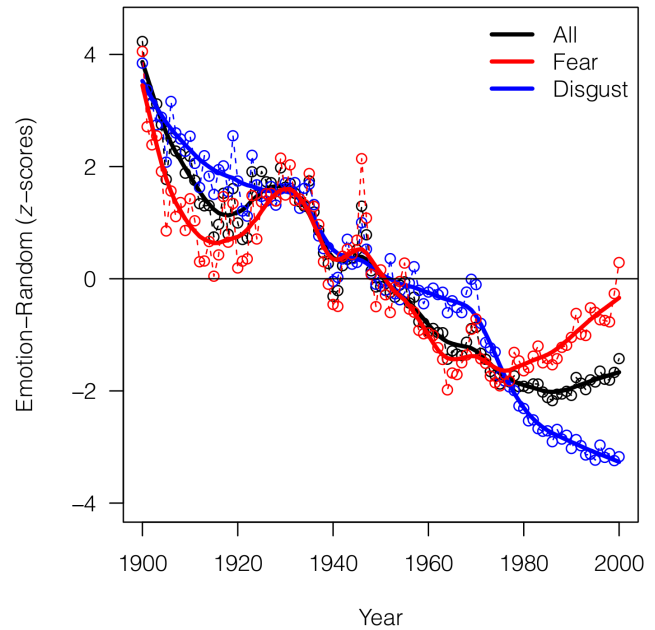


Figure 2. Decrease in the use of emotion-related words through time. Difference between *z*-scores of the six emotions and of a random sample of stems (see Methods) for years from 1900 to 2000 (raw data and smoothed trend). Red: the trend for Fear (raw data and smoothed trend), the emotion with the highest final value. Blue: the trend for Disgust (raw data and smoothed trend), the emotion with the lowest final value. Values are smoothed using Friedman’s ‘super smoother’ through R function `supsmu()` [47]. doi:10.1371/journal.pone.0059030.g002

literary emotion, we found a similar decrease in the overall use of mood words (see Figure S1).

Within this general decrease, we identify Disgust as the emotion with the lowest final *z*-score and Fear as having the highest final *z*-score (Figure 2). Notably, the mood of Fear, which declined throughout most of the early century, has increased markedly since the 1970’s, in contrast to the continued decline of other moods (Figure 2).

Our third finding is that, since about 1960, American books have increased their mood contents compared to British books. This divergence between American and British English occurs within the context of the overall decline in the use of mood words. If we plot the difference in *z*-scores between American and British word data (Figure 3a), we see a clear, steady, relative increase in American emotion-related words from 1960 to 2000. Since about 1980, books written in American have been more ‘emotional’ (in all mood figures) than the ones written in British (Figure 3a). This difference in *z*-scores – which reflects the respective deviations from each nation’s mean value – is duplicated also by the same change in absolute emotion scores (see Methods): American and British English have similar absolute emotion scores in the first half of the 20th century (or even British slightly more emotional), followed by a relative increase in the emotion scores for just the American English data set (data not shown).

Again to confirm the reality of the pattern, we checked this American–British divergence in emotion words against other indicators. Figure 3b shows the difference between American and British usage of the 307 content-free words compiled by Hughes et al. [7]. The post-1960 divergence in content-free word usage (Figure 3b) is strikingly similar to the divergence in emotional word usage (Figure 3a). As mentioned in the Introduction, changes in

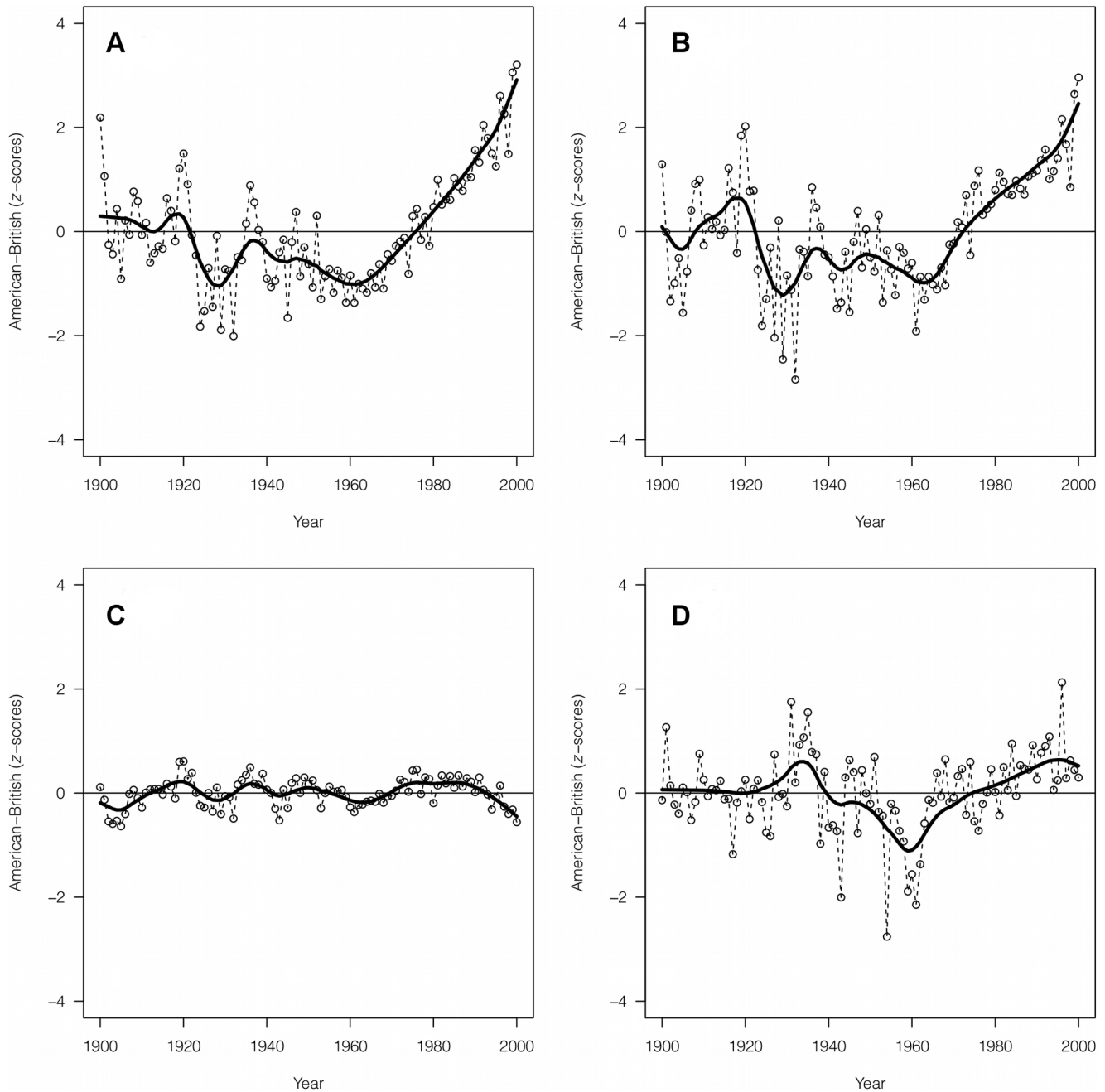


Figure 3. Differences between American English and British English. Difference between z -scores in American English and British English for years from 1900 to 2000 (raw data and smoothed trend). A: Emotion terms. B: Content-free words. C: Random sample. D: 100 largest urban agglomerations in the world. Values are smoothed using Friedman's 'super smoother' through R function `supsmu()` [47]. doi:10.1371/journal.pone.0059030.g003

the use of content-free words have been associated with broader stylistic differentiations [7], hence our analysis suggests that the divergence between American and British English in respect to the use of mood words is paired by a more general stylistic divergence since the 1960s.

As above, this does not appear to be an artifact of the Ngram data. To test this, we also checked the divergence using a random sample of words (Figure 3c) or in the usage of the names of the 100 most populated cities (Figure 3d) from the same American and British Ngram datasets. Neither showed the same change we see in both emotion words and content-free words. This suggests that

emotional content and style were coupled in a distinct way, and that the divergence of these between British and American English stands out from the background of other word types.

Discussion

Using the extraordinary new data on word frequencies in books [2], we find that significant changes in the usage of more generalized mood terms are also detectable through the years. While studies of online social media have shown how short-term patterns in word usage respond to socio-political events [9,24–26],

here we find that the expression of moods in books also reflects much longer-term trends of years or even decades. These changes in literary mood are seemingly driven by major 20th century phenomena such as World War II, The Great Depression, or the Baby Boom.

We also found a general decrease in emotional word usage in the past decades up to the present, which was observed also in fiction writing on its own. We interpret this as a genuine decrease in the literary expression of emotion, but an alternative explanation could be that mood words have changed, rather than decreased in usage, through the 20th century. This seems unlikely to explain the observed decrease, however, because we used contemporary word lists, analyzed recently in Twitter data to characterize recent events [24], any bias of which should have increased in usage towards the present.

Our results also support the popular notion that American authors express more emotion than the British. Somewhat surprisingly, this difference has apparently developed only since the 1960s, and as part of a more general stylistic differentiation in American versus British English, reflected similarly in content-free word frequencies. This relative increase of American mood word use roughly coincides with the increase of anti-social and narcissistic sentiments in U.S. popular song lyrics from 1980 to 2007 [6], as evidenced by steady increases in angry/antisocial lyrics and in the percentage of first-person singular pronouns (e.g., *I, me, mine*), with a corresponding decrease in words indicating social interactions (e.g., *mate, talk, child*) over the same 27-year period [6].

As these findings appear to genuinely reflect changes in published language, a remaining question is whether word usage represents real behavior in a population, or possibly an absence of that behavior which is increasingly played out via literary fiction (or online discourse). It has been suggested, for example, that it was the suppression of desire in ordinary Elizabethan English life that increased demand for writing “obsessed with romance and sex” [27]. So while it is easy to conclude that Americans have themselves become more ‘emotional’ over the past several decades, perhaps songs and books may not reflect the real population any more than catwalk models reflect the average body; the observed changes reflect the book market, rather than a direct change in American culture. We believe the changes do reflect changes in culture, however, because unlike lyrics of the top 10 songs, the book data are independent of book sales [2]. Although authors may not be a perfectly representative subset of the general population, at least the Google dataset is not as overtly commercial as song lyrics or any of the other ubiquitous “most popular” lists of online media. Furthermore, the association of mood changes with major 20th century economic and political events supports the fact that word usage, as retrieved from Google dataset, reveals the long term response to these events in a much broader population of book authors. The dynamics of the feedback between book authors and the wider public can be explored by future studies involving the Ngram dataset.

In any case, changes in culture consist of changes in cultural artifacts, of which words are an informative sample [2,6–8,28–30]. Future studies will surely explore diversity more closely. A population-level mean – including what we have reported here – does not necessarily track a typical behavior, so the meaning of patterns will become refined by addressing changes cross-culturally (e.g. non-English and non-Western languages), and at the smaller community scale [31]. Another promising development is the analysis of more complex sets of cultural traits that might be more diagnostic than mood words or content-free words.

More generally, we hope that we can contribute to the world of Big Data studies by showing that time depth is a crucial dimension. Our results on the long-term, mass scale encourage the more detailed use of word data to characterize the evolution of cultural differences and trends, to detect patterns previously unknown through conventional history [7,25]. While new theoretical and modelling approaches have rapidly multiplied in the field of cultural evolution (see e.g. [32–36]), we believe that the current availability and abundance of quantitative data represents an extraordinary, and much needed, opportunity to provide empirical validation in human cultural dynamics studies.

Methods

For this study we assessed the emotional valence of the text in books using a text analysis tool, namely WordNet Affect [37–39]. WordNet Affect builds on WordNet [40] by labeling synonymous terms which may represent mood states. Six mood categories, each represented by a different number of terms, have been analyzed: Anger (N = 146), Disgust (N = 30), Fear (N = 92), Joy (N = 224), Sadness (N = 115), and Surprise (N = 41). The text analysis was performed on word stems; the latter were formed using Porter’s Algorithm [41]. Both WordNet Affect and Porter’s Algorithm are considered as standard tools in text mining and have been applied in several relevant tasks [24,42–46]. We obtained the time series of stemmed word frequencies via Google’s Ngram tool (<http://books.google.com/ngrams/datasets>) in four distinct data sets: 1-grams English (combining both British and American English), 1-grams English Fiction (containing only fiction books), 1-grams American English, and 1-grams British English.

For each stemmed word we collected the amount of occurrences (case insensitive) in each year from 1900 to 2000 (both included). We excluded years before 1900 since the number of books before 1900 is considerably lower, and years after 2000 since books published recently are still being included in the data set, and therefore latest records are incomplete and possibly biased. Because the number of books scanned in the data set varies from year to year, to obtain frequencies for performing the analysis we normalized the yearly amount of occurrences using the occurrences, for each year, of the word “the”, which is considered as a reliable indicator of the total number of words in the data set. We preferred to normalize by the word “the”, rather than by the total number of words, to avoid the effect of the influx of data, special characters, etc. that may have come into books recently. The word “the” is about 5–6% of all words, and a good representative of real writing, and real sentences. To test the robustness of the normalization, we also performed the same analysis reported in Figure 1 (differences between z-scores (see below) for Joy and Sadness in the 1-grams English data set) using two alternative normalizations, namely the cumulative count of the top 10 most frequent words each year (Figure S2a), and the total counts of 1-grams as in [2] (Figure S2b). The resulting time series are highly correlated (see the legend of Figure S2), confirming the robustness of the normalization.

For a year Y , given the count C_{the} of the word “the” in the corpus as well as the counts $\{c_1, \dots, c_n\}$ of the n WordNet terms representing one of the six considered mood types, we computed a mood score (\mathcal{M}_Y) as follows:

$$\mathcal{M}_Y = \frac{1}{n} \sum_{i=1}^n \frac{c_i}{C_{\text{the}}}, \quad (1)$$

i.e. a mood score is essentially the average normalized frequency across the considered mood terms. In order to compare different

types of moods effectively, after computing the mood scores for the entire set of years (1900 to 2000), we converted them to their z -score equivalent ($\mathcal{M}z_Y$), using:

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}}, \quad (2)$$

where $\mu_{\mathcal{M}}$ and $\sigma_{\mathcal{M}}$ denote the mean and standard deviation of the mood scores across the considered set of years.

When determining the ‘absolute’ trends of moods (results in Figure 2), to avoid any possible bias related to normalization, we compared the z -scores of moods time series with a z -score derived by a random sample of stems. This random sample was obtained by extracting 10,000 random terms from the Part of Speech database, which is a combination of Moby Part-of-Speech II and WordNet database, and contains 295,172 terms (<http://wordlist.sourceforge.net/pos-readme>). The terms were stemmed using Porter’s Algorithm [41]. Duplicated stems and stems shorter than two letters were finally eliminated obtaining a sample of $N = 8967$ stems that we used for the analysis.

Finally (results in Figure 3), we compared the difference between American English and British English books in three other data sets (in addition to mood terms): content-free words, using the list of the 307 content-free words provided by Huges et al. [7], the same random stems used for Figure 2, and the list of the 100 largest urban agglomerations in the world (from [http://en.wikipedia.org/wiki/List_of_urban_agglomerations_by_population_\(United_Nations\)](http://en.wikipedia.org/wiki/List_of_urban_agglomerations_by_population_(United_Nations))), containing terms like *Tokyo*, *Karachi*, and *Berlin*. Agglomerations with composite names, e.g. *New York–Newark*, *Rio de Janeiro* were excluded).

References

- Weinberger D (2011) The machine that would predict the future. *Scientific American* 305(6): 52–57.
- Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–182.
- Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) Quantifying the evolutionary dynamics of language. *Nature* 449: 713–716.
- Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449: 717–720.
- Bentley RA (2008) Random drift versus selection in academic vocabulary. *PLoS ONE* 3(8): e3057.
- DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics. *Psychology of Aesthetics, Creativity and the Arts* 5: 200–207.
- Hughes JM, Foti NJ, Krakauer DC, Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences USA* 109: 7682–7686.
- Twenge J, Campbell KW, Gentile B (2012) Increases in individualistic words and phrases in American books, 1960–2008. *PLoS ONE* 7(7): e40181.
- Bollen J, Mao H, Zeng X-J (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2: 1–8.
- Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford: 1–12.
- Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*: 1320–1326.
- Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of Twitter data. *Proceedings of the Workshop on Languages in Social Media*: 30–38.
- Altmann EG, Pierrehumbert JB, Motter AE (2011) Niche as a determinant of word fate in online groups. *PLoS ONE* 6(5): e19009.
- Bentley RA, Ormerod P (2010) A rapid method for assessing social versus independent interest in health issues. *Social Science and Medicine* 71: 482–485.
- Bentley RA, Garnett P, O’Brien MJ, Brock WA (2012) Word Diffusion and Climate Science. *PLoS ONE*, 7(11): e47966.
- Saavedra S, Hagerty K, Uzzi B (2011) Synchronicity, instant messaging, and performance among financial traders. *Proceedings of the National Academy of Sciences USA* 108: 5296–5301.
- Saavedra S, Duch J, Uzzi B (2011) Tracking Traders’ Understanding of the Market Using e-Communication Data. *PLoS ONE* 6(10): e26705.
- Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. *Scientific Reports* 2: Article 350.
- Fortunato L, Jordan F (2010) Your place or mine? A phylogenetic comparative analysis of marital residence in Indo-European and Austronesian societies. *Philosophical Transactions Royal Society B* 365: 3913–3922.
- Evans N, Levinson SC (2009) The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492.
- Nettle D (1999) Linguistic diversity of the Americas can be reconciled with a recent colonisation. *Proceedings of the National Academy of Sciences USA* 96: 3325–3329.
- Nettle D (2007) Language and genes: A new perspective on the origins of human cultural diversity. *Proceedings of the National Academy of Sciences USA* 104: 10755–10756.
- Nichols J (2008) Language spread rates and prehistoric American migration rates. *Current Anthropology* 49: 1109–1118.
- Lansdall T, Lamos V, Cristianini N (2012) Effects of the Recession on Public Mood in the UK. *Proceedings of WWW’12*: 1221–1226.
- O’Connor B, Balasubramanian B, Routledge BR, Smith NA (2012). From tweets to polls: Linking sentiment to public opinion time series. *Proceedings of the International AAI Conference on Weblogs and Social Media*. Menlo Park, CA: The AAAI Press.
- Schellenberg EG, von Scheve C (2012) Emotional cues in American popular music: Five decades of the Top 40. *Psychology of Aesthetics, Creativity, and the Arts* 6: 196–203.
- Shapiro J (2012) The Queen’s revels. *New York Times Book Review* (15 July 2012): 14.
- Hahn MW, Bentley RA (2003) Drift as a mechanism for cultural change: An example from baby names. *Proceedings of the Royal Society B* 270: S1–S4.

Supporting Information

Figure S1 Decrease in the use of emotion-related words through time in fiction books. Difference between z -scores of the six emotions and of a random sample of stems (see Methods) for years from 1900 to 2000 (raw data and smoothed trend) in the 1-grams English Fiction data set. Red: the trend for Fear (raw data and smoothed trend), the emotion with the highest final value. Blue: the trend for Disgust (raw data and smoothed trend), the emotion with the lowest final value. Values are smoothed using Friedman’s ‘super smoother’ through R function `supsmu()` [47]. (TIFF)

Figure S2 Historical periods of positive and negative moods with alternative normalizations. Difference between z -scores of Joy and Sadness for years from 1900 to 2000 (raw data and smoothed trend). Values above zero indicate generally ‘happy’ periods, and values below the zero indicate generally ‘sad’ periods. Values are smoothed using Friedman’s ‘super smoother’ through R function `supsmu()` [47]. A: Frequencies are normalized using the cumulative count of the top 10 most frequent words for each year. Correlations with the time series used in the analysis (normalized with the yearly count of “the”) are statistically significant (Pearson’s $r = 0.97$ for row data, and $r = 0.96$ for smoothed data. In both cases $p < 10^{-10}$ and $N = 101$). B: Frequencies are normalized using the total counts of 1-grams for each year. Also in this case correlations with the time series used in the analysis are statistically significant (Pearson’s $r = 0.91$ for row data, and $r = 0.86$ for smoothed data. In both cases $p < 10^{-10}$ and $N = 101$). (TIFF)

Author Contributions

Conceived and designed the experiments: AA VL RAB. Analyzed the data: AA VL. Wrote the paper: AA VL PG RAB.

29. Berger J, Le Mens G (2009) How adoption speed affects the abandonment of cultural tastes. *Proceedings of the National Academy of Sciences USA* 106: 8146–8150.
30. Acerbi A, Ghirlanda S, Enquist M (2012) The Logic of Fashion Cycles. *PLoS ONE* 7(3): e32541.
31. Mateos P, Longley PA, O'Sullivan D (2011) Ethnicity and Population Structure in Personal Naming Networks. *PLoS ONE* 6(9): e22943.
32. Sperber D (1996) *Explaining culture. A naturalistic approach.* Oxford: Blackwell.
33. Bentley RA, Hahn MW, Shennan SJ (2004) Random drift and culture change. *Proceedings of the Royal Society B* 271: 1443–1450.
34. Mesoudi A, Whiten A, Laland KN (2006) Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* 29(4): 329–383.
35. Acerbi A, Enquist M, Ghirlanda S (2009) Cultural evolution and individual development of openness and conservatism. *Proceedings of the National Academy of Sciences USA* 106(45): 18931–18935.
36. Rendell L, Boyd R, Cownden D, Enquist M, Eriksson K, et al. (2010) Why copy others? insights from the social learning strategies tournament. *Science* 328: 208–213.
37. Strapparava C, Valitutti A (2004) Wordnet-affect: an affective extension of wordnet. *Proceedings of the 4th International Conference on Language Resources and Evaluation*: 1413–1418.
38. Strapparava C, Valitutti A, Stock O (2006) The affective weight of lexicon. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*: 423–426.
39. Strapparava C, Mihalcea R (2008) Learning to identify emotions in text. *Proceedings of SAC'08*: 1556–60.
40. Miller GA (1995) WordNet: a lexical database for English. *CACM* 38(11): 39–41.
41. Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3): 130–137.
42. Chaumartin FR (2007) UPAR7: A knowledge-based system for headline sentiment tag-ging. *Proceedings of the 4th International Workshop on Semantic Evaluations.*, 422–425.
43. Neviarouskaya A, Prendinger H, Ishizuka M (2007) Textual affect sensing for sociable and expressive online communication. *Affective Computing and Intelligent Interaction*: 218–229.
44. Hu X, Downie JS, Ehmann AF (2009) Lyric text mining in music mood classification. *10th International Society for Music Information Retrieval Conference*: 411–416.
45. Calvo RA, D'Mello S (2010) Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *TAC* 1(1): 18–37.
46. Roberts K, Harabagiu SM (2012) Statistical and Similarity Methods for Classifying Emotion in Suicide Notes. *Biomedical Informatics Insights (Suppl. 1)*: 195–204.
47. R Development Core Team (2010) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna: The R Project for Statistical Computing. Available: <http://www.R-project.org>. Accessed 2013 Feb 22.