Using human evaluation of 100,000 words spread across 24 corpora in 10 languages diverse in origin and culture, we present evidence of a deep imprint of human sociality in language, observing that (i) the words of natural human language possess a universal positivity bias, (ii) the estimated emotional content of words is consistent between languages under translation, and (iii) this positivity bias is strongly independent of frequency of word use. Alongside these general regularities, we describe interlanguage variations in the emotional spectrum of languages that allow us to rank corpora. We also show how our word evaluations can be used to construct physical-like instruments for both real-time and offline measurement of the emotional content of large-scale texts.

Human language, our great social technology, reflects that which it describes through the stories it allows to be told and us, the tellers of those stories. Although language’s shaping effect on thinking has long been controversial (1–3), we know that a rich array of metaphor encodes our conceptualizations (4), which it describes through the stories it allows to be told and us, the tellers of those stories. Although language’s shaping effect on thinking has long been controversial (1–3), we know that a rich array of metaphor encodes our conceptualizations (4), and (5) the estimated emotional content of words is consistent between languages under translation, and (iii) this positivity bias is strongly independent of frequency of word use. Alongside these general regularities, we describe interlanguage variations in the emotional spectrum of languages that allow us to rank corpora. We also show how our word evaluations can be used to construct physical-like instruments for both real-time and offline measurement of the emotional content of large-scale texts.

Materials and Methods

To explore the positivity of human language deeply, we constructed 24 corpora spread across 10 languages. Our global coverage of linguistically and culturally diverse languages includes English, Spanish, French, German, Brazilian Portuguese, Korean, Chinese (Simplified), Russian, Indonesian, and Arabic. The sources of our corpora are similarly broad, spanning books (14), news outlets, social media, the web (15), television and movie subtitles, and music lyrics (16). Our work here greatly expands upon our earlier study of English alone, where we found strong evidence for a use-invariant positivity bias (17). In SI Appendix, we provide full details of our corpora (SI Appendix, Table S1), survey, and participants (SI Appendix, Table S2).

We address the social nature of language in two important ways: (i) We focus on the words people most commonly use, and (ii) we measure how those same words are received by individuals. We take word use frequency as the primary organizing measure of a word’s importance. Such a data-driven approach is crucial for both understanding the structure of language and creating linguistic instruments for principled measurements (18, 19). By contrast, earlier studies focusing on meaning and emotion used “expert” generated word lists, and these word lists fail statistically to match frequency distributions of natural language (10–12, 20), confounding attempts to make claims about language in general. For each of our corpora, we selected between 5,000 and 10,000 of the most frequently used words, choosing the exact numbers so that we obtained ~10,000 words for each language.

Of our 24 corpora, we received 17 already parsed into words by the source: the Google Books Project (six corpora), the GoogLe Web Crawl (eight corpora), and movie and television subtitles (three corpora). For the other seven corpora (five Twitter corpora, the New York Times, and music lyrics), we extracted words by standard white space separation. Twitter was easily the most variable and complex of our text sources, and required additional treatment. In parsing Twitter, we required strings to contain at least one Unicode character and no invisible control characters, and we excluded strings representing web links, bearing a leading @, ampersand (&), or other punctuation (e.g., Twitter IDs) but kept hashtags. Finally, for all corpora, we converted words to lowercase. We observed that common English words appeared in the Twitter corpora of other languages, and we have chosen simply to acknowledge this reality of language and allow these commonly used borrowed words to be evaluated.

Although there are many complications with inflections and variable orthography, we have found merit for our broad analysis in not collapsing related words. For example, we have observed that allowing different conjugations of verbs to stand in our corpora is valuable because human evaluations of such have proved to be distinguishable (e.g., present vs. past tense (18)). As should be expected, a more nuanced treatment going beyond the present paper’s bounds by involving stemming and word type, for example, may lead to minor corrections (21), although our central observations will remain robust and in no way change the behavior of the instruments we generate.

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 the most commonly used words in 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 $s$ such that the union of all words with a rank $\leq r$ in at least one corpus formed a set of at least 10,000 words.
We then paid native speakers to rate how they felt in response to individual words on a nine-point scale, with 1 corresponding to most negative or saddest, 5 to neutral, and 9 to most positive or happiest (10, 18) (SI Appendix). This happy–sad semantic differential (20) functions as a coupling of two standard five-point Likert scales. Participants were restricted to certain regions or countries (e.g., Portuguese was rated by residents of Brazil). Overall, we collected 50 ratings per word for a total of around 5 million individual human assessments. We provide all datasets as part of SI Appendix.

**Results and Discussion**

In Fig. 1, we show distributions of the average happiness scores for all 24 corpora, leading to our most general observation of a clear positivity bias in natural language. We indicate the above-neutral part of each distribution with yellow and the below-neutral part with blue, and we order the distributions moving upward by increasing median (vertical red line). For all corpora, the median clearly exceeds the neutral score of 5. The background gray lines connect deciles for each distribution. In SI Appendix, Fig. S1, we provide the same distributions ordered instead by increasing variance.

As is evident from the ordering in Fig. 1 and SI Appendix, Fig. S1, although a positivity bias is the universal rule, there are minor differences between the happiness distributions of languages. For example, Latin American-evaluated corpora (Mexican Spanish and Brazilian Portuguese) exhibit relatively high medians and, to a lesser degree, higher variances. For other languages, we see that those languages with multiple corpora have more variable medians, and specific corpora are not ordered by median in the same way across languages (e.g., Google Books has a lower median than Twitter for Russian, but the reverse is true for German and English). In terms of emotional variance, all four English corpora are among the highest, whereas Chinese and Russian Google Books seem especially constrained.

We now examine how individual words themselves vary in their average happiness score between languages. Owing to the scale of our corpora, we were compelled to use an online service, choosing Google Translate. For each of the 45 language pairs, we translated isolated words from one language to the other and then back. We then found all word pairs that (i) were translationally stable, meaning the forward and back translation returns the original word, and (ii) appeared in our corpora for each language.

We provide the resulting comparison between languages at the level of individual words in Fig. 2. We use the mean of each language’s word happiness distribution derived from its merged corpora to generate a rough overall ordering, acknowledging that frequency of use is no longer meaningful, and moreover is not relevant, because we are now investigating the properties of individual words. Each cell shows a heat map comparison with word density increasing as shading moves from gray to white. The background colors reflect the ordering of each pair of languages: yellow if the row language had a higher average happiness than the column language and blue for the reverse. Also, in each cell, we display the number of translation-stable words between language pairs, N, along with the difference in average word happiness, Δ, where each word is equally weighted.

A linear relationship is clear for each language–language comparison, and is supported by Pearson’s correlation coefficient r being in the range 0.73–0.89 (P < 10⁻¹¹⁸ across all pairs; Fig. 2 and SI Appendix, Tables S3–S5). Overall, this strong agreement between languages suggests that approximate estimates of word happiness for un-scored languages could be generated with no expense from our existing dataset. Some words will, of course, translate unsatisfactorily, with the dominant meaning changing between languages. For example “lying” in English, most readily interpreted as speaking falsehoods by our participants, translates to “acostado” in Spanish, meaning recumbent. Nevertheless, happiness scores obtained by translation will be serviceable for purposes where the effects of many different words are incorporated.

![Fig. 1. Distributions of perceived average word happiness h_{avg} for 24 corpora in 10 languages. The histograms represent the 5,000 most commonly used words in each corpora (details are provided in SI Appendix), and native speakers scored words on a 1–9 double-Likert scale, with 1 being extremely negative, 5 being neutral, and 9 being extremely positive. Yellow indicates positivity (h_{avg} > 5), and blue indicates negativity (h_{avg} < 5). Distributions are ordered by increasing median (red vertical line). The background gray lines connect deciles of adjacent distributions. The same distributions arranged according to increasing variance are shown in SI Appendix, Fig. S1.](http://www.pnas.org/cgi/doi/10.1073/pnas.1411678112)
Stepping back from examining interlanguage robustness, we return to a more detailed exploration of the rich structure of each corpus’ happiness distribution. In Fig. 3, we show how average word happiness $h_{avg}$ is largely independent of word use frequency for four example corpora. We first plot use frequency rank $r$ of the 5,000 most frequently used words as a function of their average happiness score, $h_{avg}$ (background dots), along with some example evenly spaced words. We note that words at the extremes of the happiness scale are ones evaluators agreed upon strongly, whereas words near neutral range from being clearly neutral [e.g., $h_{avg}$ (“the”) = 4.98] to contentious with high SD (17). We then compute deciles for contiguous sets of 500 words, sliding this window through rank $r$. These deciles form the vertical strands. We overlay randomly chosen, equally spaced example words to give a sense of each corpus’ emotional texture.

We chose the four example corpora shown in Fig. 3 to be disparate in nature, covering diverse languages (French, Egyptian Arabic, Brazilian Portuguese, and Chinese), regions of the

**Fig. 2.** Scatter plots of average happiness for words measured in different languages. We order languages from relatively most positive (Spanish) to relatively least positive (Chinese); a yellow background indicates the row language is more positive than the column language, and a blue background indicates the converse. The overall plot matrix is symmetrical about the leading diagonal, with the redundancy allowing for easier comparison between languages. In each scatter plot, the key gives the number of translation-stable words for each language pair, $N$; the average difference in translation-stable word happiness between the row language and the column language, $\Delta$; and the Pearson correlation coefficient for the regression, $r$. All $P$ values are less than $10^{-118}$ for the Pearson correlation coefficient and less than $10^{-82}$ for the Spearman correlation coefficient.
world (Europe, the Middle East, South America, and Asia), and
texts [Twitter, movies and television, the web (15), and books
(14)]. The remaining 20 corpora yield similar plots (SI Appendix),
and all corpora also exhibit an approximate self-similarity in SD
for word happiness.

Across all corpora, we observe visually that the deciles tend
to stay fixed or move slightly toward the negative, with some
expected fragility at the 10% and 90% levels (due to the dis-
tributions’ tails), indicating that the overall happiness distri-
bution of each corpus approximately holds independent of
word use. In Fig. 3, for example, we see that both the
Brazilian Portuguese and French examples show a small shift to
the negative for increasingly rare words, whereas there is no
visually clear trend for the Arabic and Chinese cases. Fitting
$h_{avg} = \alpha r + \beta$ typically returns $\alpha$ on the order of $-1 \times 10^{-5}$
suggesting $h_{avg}$ decreases 0.1 per 10,000 words. For SDs of
happiness scores, we find a similarly weak drift toward higher
values for increasingly rare words (correlations and linear fits
for $h_{avg}$ and $h_{std}$ as a function of word rank $r$ for all corpora
are provided in SI Appendix, Tables S6 and S7). We thus find that,
to first order, not just the positivity bias but the happiness
distribution itself applies for common words and rare words
alike, revealing an unexpected addition to the many well-known
scalings found in natural language, as famously exemplified by
Zipf’s law (22).

In constructing language-based instruments for measuring ex-
pressed happiness, such as our hedonometer (18), this frequency
independence allows for a way to “increase the gain” in a fashion
resembling standard physical instruments. Moreover, we have ear-
lier demonstrated the robustness of our hedonometer for the

![Fig. 3. Examples of how word happiness varies little with use frequency.](https://www.pnas.org/cgi/doi/10.1073/pnas.1411678112)
English language, showing, for example, that measurements derived from Twitter correlate strongly with Gallup well-being polls and related indices at the state and city level for the United States (19).

Here, we provide an illustrative use of our hedonometer in the realm of literature, inspired by Vonnegut’s shapes of stories (23, 24). In Fig. 4, we show “happiness time series” for three famous works of literature, evaluated in their original languages of English, Russian, and French, respectively: Melville’s Moby Dick (www.gutenberg.org), Dostoyevsky’s Crime and Punishment (25), and Dumas’ The Count of Monte Cristo (www.gutenberg.org). We slide a 10,000-word window through each work, computing the average happiness using a “lens” for the hedonometer in the following manner. We capitalize on our instrument’s tunability to obtain a strong signal by excluding all words for which $3 < h_{\text{avg}} < 7$ (i.e., we keep words residing in the tails of each distribution) (18). Denoting a given lens by its corresponding set of allowed words $L$, we estimate the happiness score of any text $T$ as $h_{\text{avg}}(T) = \frac{\sum_{w \in T} f_w h_{\text{avg}}(w)}{\sum_{w \in T} f_w}$, where $f_w$ is the frequency of word $w$ in $T$ (16).

The three resulting happiness time series provide interesting detailed views of each work’s narrative trajectory revealing

![Fig. 4. Emotional time series for three famous 19th century works of literature: Melville’s Moby Dick (Top), Dostoyevsky’s Crime and Punishment (Middle), and Dumas’ The Count of Monte Cristo (Bottom). Each point represents the language-specific happiness score for a window of 10,000 words (converted to lowercase), with the window translated throughout the work. The overlaid word shifts (A–I) show example comparisons between different sections of each work. Word shifts indicate which words contribute the most toward and against the change in average happiness between two texts (SI Appendix, pp. S9–S10). Although a robust instrument in general, we acknowledge the hedonometer’s potential failure for individual words due to both language evolution and words possessing more than one meaning. For Moby Dick, we excluded “cried” and “cry” (to speak loudly rather than weep) and “Coffin” (surname, still common in Nantucket). Such alterations, which can be done on a case-by-case basis, do not noticeably change the overall happiness curves, although leaving the word shifts more informative. We provide online, interactive visualizations of the emotional trajectories of over 10,000 books at hedonometer.org/books.html.](https://www.gutenberg.org)
numerous peaks and troughs throughout, at times clearly dropping below neutral. Both Moby Dick and Crime and Punishment end on low notes, whereas The Count of Monte Cristo culminates with a rise in positivity, accurately reflecting the finishing arcs of all three. The “word shifts” overlaying the time series compare two distinct regions of each work, showing how changes in word abundances lead to overall shifts in average happiness. Such word shifts are essential tests of any sentiment measurement, and are made possible by the linear form of our instrument (16, 18) [a full explanation is provided in SI Appendix, pp. S9–S10]. As one example, the third word shift for Moby Dick shows why the average happiness of the last 10% of the book is well below the average happiness of the first 25%. The major contribution is an increase in relatively negative words, including “missing,” “shot,” “poor,” “die,” and “evil.”

By adjusting the lens, many other related time series can be formed, such as those produced by focusing on only positive or negative words. Emotional variance as a function of text position can also be readily extracted. At hedonometer.org/books.html, we provide online, interactive emotional trajectories for over 10,000 works of literature where different lenses and regions of comparisons may be easily explored. Beyond this example tool we have created here for the digital humanities and our hedonometer for measuring population well-being, the datasets we have generated for the present study may be useful in creating a great variety of language-based instruments for assessing emotional expression.

Overall, our major scientific finding is that when experienced in isolation and weighted properly according to use, words, which are the atoms of human language, present an emotional spectrum with a universal, self-similar positive bias. We emphasize that this apparent linguistic encoding of our social nature is a system-level property, and in no way asserts all natural texts will skew positive (as exemplified by certain passages of the three works in Fig. 4) or diminishes the salience of negative states (26). Going forward, our word happiness assessments should be periodically repeated and carried out for new languages, tested on different demographics, and expanded to phrases both for the improvement of hedonometric instruments and to chart the dynamics of our collective social self.

ACKNOWLEDGMENTS. We thank M. Shields, K. Corner, N. Berry, I. Ramiscal, C. Burke, P. Carrigan, M. Koehler, and Z. Henschel, in part, for their roles in developing hedonometer.org. We also thank F. Henegan, A. Powers, and N. Atkins for conversations. P.S.D. was supported by National Science Foundation CAREER Award 0846668.