# Positive words carry less information than negative words

**David Garcia, Antonios Garas, and Frank Schweitzer**
Chair of Systems Design, ETH Zurich, Kreuzplatz 5, 8032 Zurich, Switzerland

**Abstract**

We show that the frequency of word use is not only determined by the word length [1] and the average information content [2], but also by its emotional content. We have analysed three established lexica of affective word usage in English, German, and Spanish, to verify that these lexica have a neutral, unbiased, emotional content. Taking into account the frequency of word usage, we find that words with a positive emotional content are more frequently used. This lends support to Pollyanna hypothesis [3] that there should be a positive bias in human expression. We also find that negative words contain more information than positive words, as the informativeness of a word increases uniformly with its valence decrease. Our findings support earlier conjectures about (i) the relation between word frequency and information content, and (ii) the impact of positive emotions on communication and social links.

Keywords: *communication; emotion; language; information theory.*

## 1 Introduction

One would argue that human languages, to support an efficient communication, are not biased towards positive or negative emotions. We have tested this argument by analysing established lexica of affective word usage of three different languages, namely English [4], German [5], and Spanish [6]. We find that the emotional content averaged over all words in these lexica is indeed neutral. Considering, however, the everyday usage frequency of these words we find that the overall emotion of the three languages is strongly biased towards positive values, because words with a positive emotion are more frequently used than those with a negative emotion. These results are consistent with previous works on the relation between emotion and word frequency [7, 8] for English in corpora of limited size.

Historically, the *frequency* of words was first analysed as a function of the word *length*. This dependency is usually described as Zipf's law, following the pioneering work of Zipf [1, 9] who showed that frequency predicts the length of a word as result of a principle of least effort. Zipf's law highlighted fundamental principles of organisation in human language [10], and called for an interdisciplinary approach to understand its origin [11–13]. Recently Piantadosi et al. [2] extended Zipf's approach by showing that word lengths are determined such that the efficiency of communication in a noisy channel is maximised. In other words, word length increases with information content, in order to have efficient communication. Further discussions [14–16] highlighted the relevance of *meaning* as part of the communication process as, for example, more abstract ideas are expressed through longer words [17]. Our work focuses on one particular aspect of meaning, namely the *emotion* expressed in a word. This requires additional data beyond word frequency, which became available thanks to large datasets of human behaviour on the Internet. Millions of individuals write text online, for which a quantitative analysis can provide new insights into the structure of human language and even provide a validation of social theories [18].

Our investigation combines two analyses (see Materials and Methods), (i) quantifying the emotional content of words in terms of valence, and (ii) quantifying the frequency of word usage. In order to link the emotionality of each word with the information it carries, we build on the recent work of Piantadosi et al. [2]. This way, we reveal the existence of an emotional dimension in the communication process which influences the information carried by words. While the rational process that optimises communication determines word lengths by the information they carry [2], we find that the emotional content affects the word frequency such that positive words appear more frequently. This points towards an emotional bias in used language and supports Pollyanna hypothesis [3], which asserts that there is a bias towards the usage of positive words. The relation between word frequency and information content, on the other hand, leads to the conclusion that positive, i.e. more frequent, words carry less information than negative ones. In other words, the informativeness of words highly depends on their emotional polarity.

## 2 Results

In detail, we have analysed three lexica of affective word usage which contain 1034 English words, 2902 German words and 1034 Spanish words, together with their emotional scores obtained from extensive human ratings. These lexica have effectively established the standard for emotion analyses of human texts [20]. Each word in these lexica is assigned a set of values measuring different aspects of word emotionality. One of these values, a scalar variable $v$ called valence [21], represents the degree of pleasure induced by the emotion associated with the word. In this paper, we use $v$ to quantify word emotionality.

In each lexicon, words were chosen such that they evenly span the full range of valence. [*] In order to compare the emotional content of the three different languages, we have rescaled all values of $v$ to the interval [-1,1]. As shown in the left panel of Fig. 1, indeed, the average valence, as well as the median, of all three lexica is very close to zero, i.e. they do not provide an emotional bias.

This analysis, however, neglects the actual frequency of word usage, which is highly skew distributed [1, 9]. Instead of applying Zipf's law to calculate the actual appearance of a word, we have used Google's $N$-gram dataset (see Materials and Methods) which, with $10^{12}$ tokens, is

[*] The lexica focus on single words rather than on phrases or longer expressions.

one of the largest datasets available about real human text expressions in the Internet. For our analysis, we have chosen those words which have an affective classification in the respective lexicon in either English, German, or Spanish. The different usage of words with the same valence is quite obvious. For example, both words "party" and "sunrise" have the same positive valence of 0.715, however the frequency of "party" is 144.7 per one million words compared to 6.8 for "sunrise". Similarly, both "dead" and "distressed" have a negative valence of -0.765, but the former appears 48.4 times per one million words, the latter only 1.6 times. Taking into account all frequencies of word usage, we find for all three languages that the median shifts considerably towards positive values. This is shown in the right panel of Fig. 1. Hence, with respect to usage we find evidence that human languages are emotionally charged, i.e. significantly different from being neutral. This affects quantitative analyses of the emotions in written text, because the "emotional reference point" is not at zero, but at considerably higher valence values (about 0.3).

Our analysis suggests that there is a definite relation between word valence and frequency of use, the latter impacting on human communication as already discussed in [2]. Here we study the role of emotions in the communication process building on the relation between information measures and valence. While we are unable to measure information perfectly, we restrict ourselves to the so-called self-information, $I(w)$, which estimates the information content from the probability of appearance of a word, $P(w)$, as $I(w) = -\log P(w)$ [22]. I.e., very common words provide less information than very unusual ones. For the three lexica, we calculated $I(w)$ of each word and linked it to its valence, $v(w)$. Fig. 2 shows the results. From the clear negative correlation found for all three languages (between -0.3 and -0.4), we deduce that words with less information content carry more positive emotions, as the average valence decreases along the self-information range. A detailed statistical analysis is provided in Table 1 in the Appendix.

Our results outperform a recent finding [7] that, while focusing on individual text production, reported a weaker correlation (below 0.3) between the logarithm of word usage frequency and valence. That analysis was based on a much smaller data set from Internet discussions (in the order of $10^8$ tokens) and the same English lexicon of affective word usage [4] we used. Using a much higher accuracy in estimating word frequencies and extending the analysis to three different languages, we were able to verify that there is a significant relation between the emotional content of a word and its self-information, impacting the frequency of usage.

In order to further strengthen our results, we tested different hypotheses impacting the relation between word usage and valence (see the Tables in the Appendix). Firstly, we tested if the self-information is just correlated with the absolute value of the valence, this way ignoring differences between positive and negative emotions. This could not be confirmed (see Table 1), which means indeed that the usage frequency of a word is not just related to the overall emotional intensity, but to the positive or negative emotion expressed. Secondly, we controlled for the sole influence of the word length, to find out that this has no significant effect on our results (see Table 1).

Thirdly, we evaluated how the context of a word impacts its informativeness. Therefore, we analysed the frequency of sequences of $N$ words, called $N$-grams, from the Google dataset for $N \in \{2, 3, 4\}$ (see Materials and Methods). Instead of the self-information that holds for single words, we used the information content from entropy estimations of the $N$-grams in the same manner as given by Piantadosi et al. [2]. As Table 2 shows, the correlation coefficients between valence and information content for $N$-grams are found to be significant and consistent with the case of single words.

Eventually, we also performed a control analysis using alternative frequency datasets, to account for possible anomalies in the Google dataset due to its online origin. We used the word frequencies estimated from traditional written corpuses, as reported in the original datasets [4–6]. Calculating the self-information from these and relating them to the valences given, we obtained similar, but slightly lower Pearson's correlation coefficients (see Table 2). So, we conclude that our results are robust across different types of written communication, for the three languages analysed.

## 3 Discussion

Our analysis provides strong evidence that words with a positive emotional content are more often used. This lends support to Pollyanna hypothesis [3], i.e. positive words are more often used, for all the three languages studied. Our conclusions are consistent for, and independent of, different corpuses used to obtain the word frequencies, i.e. they are shown to hold for traditional corpuses of formal written text, as well as for the Google dataset and cannot be attributed as artifacts of Internet communication.

Furthermore, we have pointed out the relation between the emotional and the informational content of words. Words with negative emotions are less often used. But because of their rareness they carry more self-information than positive words. This relation remains valid even when considering the context of a word in sequences up to four words ($N$-grams). Controlling for word length, we find that the correlation between information and valence does not depend on the length, i.e. it is indeed the usage frequency that matters.

In our analysis, we did not explore the role of syntactic rules and grammatical classes such as verbs, adjectives, etc. However, previous studies have shown the existence of a similar bias when studying adjectives and their negations [8]. The question of how syntax influences emotional expression is beyond the scope of the present work.

The findings reported in this paper suggest that the process of communication between humans, which is known to optimise information transfer [2], also creates a bias towards positive emotional content. A possible explanation for this is basic impact of positive emotions on the formation of social links between humans. Human communication should reinforce such links which it both shapes and depends on. So, it makes much sense that human languages on average have strong bias towards positive emotions, as we have shown (see Figure 1). Negative expressions, on the other hand, mostly serve a different purpose, namely that of transmitting highly informative and relevant events. They are less used, but carry more information.

Our findings are also consistent with emotion research in social psychology. According to [23], the expression of positive emotions increases the level of communication and strengthens social links. This would lead to stronger pro-social behaviour and cooperation, giving evolutionary advantage to societies whose communication shows a positive bias. As a consequence, positive sentences would become more frequent and even advance to a social norm (cf. “Have a nice day”), but they would provide less information when expressed. Eventually, we emphasise that the positive emotional “charge” of human communication has a further impact on the quantitative analysis of communication in the Internet, for example in chatrooms, fora, blogs, and other online communities. Our analysis provides an estimation of the emotional baseline of human written expression, and automatic tools and further analyses will need to take this into account.

## 4 Materials and Methods

We estimate the information measures used in this study from the Google $N$-gram dataset [19]. This dataset contains frequencies for single words and $N$-grams (word sequences) up to size five, calculated from an online corpus of more than a trillion tokens. The source of this dataset is the whole Google crawl, which aimed at spanning a large subset of the web, providing a wide point of view on how humans write on the Internet.

Self-information is calculated from single word frequencies as given above, and information content is extracted by taking into account neighbouring words as context. Given each context $c_i$ where a word $w$ appears, the information content is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\log(\mathrm{P}(W=w|C=c_i)) \tag{1}$$

where $N$ is the total frequency of the word in the corpus used for the estimation. These values were calculated as approximations of the context given the words preceding $w$ up to size 4 in [2].

## Acknowledgment

This research has received funding from the European Community’s Seventh Framework Programme FP7-ICT-2008-3 under grant agreement no 231323 (CYBEREMOTIONS).

## References

[1] Zipf, G. K. (1935) *The Psycho-Biology of Language.* (Houghton Mifflin, Oxford, England:), p. 336.

[2] Piantadosi, S, Tily, H, & Gibson, E. (2011) Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* **108**, 3526.

[3] Boucher, J and Osgood, C. k (1969) The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior* **8**, 1–8.

[4] Bradley, M. M & Lang, P. J. (1999) Affective norms for English words (ANEW): Instruction manual and affective ratings, (The Center for Research in Psychophysiology, University of Florida.), Technical report.

[5] Võ, M. L.-H, Conrad, M, Kuchinke, L, Urton, K, Hofmann, M. J, & Jacobs, A. M. (2009) The Berlin Affective Word List Reloaded (BAWL-R). *Behavior research methods* **41**, 534–8.

[6] Redondo, J, Fraga, I, Padrón, I, & Comesaña, M. (2007) The Spanish adaptation of ANEW (affective norms for English words). *Behavior research methods* **39**, 600–5.

[7] Augustine, A. A, Mehl, M. R, & Larsen, R. J. (2011) A Positivity Bias in Written and Spoken English and Its Moderation by Personality and Gender. *Social Psychological and Personality Science* **2**, 508–515.

[8] Rozin, P, Berman, L, & Royzman, E. (2010) Biases in use of positive and negative words across twenty natural languages. *Cognition & Emotion* **24**, 536–548.

[9] Zipf, G. K. (1949) *Human behavior and the principle of least effort.* (Addison-Wesley, New York).

[10] Ferrer i Cancho, R & Sole, R. V. (2003) Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 788–91.

[11] Hauser, M. D, Chomsky, N, & Fitch, W. T. (2002) The faculty of language: what is it, who has it, and how did it evolve? *Science (New York, N.Y.)* **298**, 1569–79.

[12] Kosmidis, K, Kalampokis, A, & Argyrakis, P. (2006) Statistical mechanical approach to human language. *Physica A: Statistical Mechanics and its Applications* **366**, 495–502.

[13] Havlin, S. (1995) The distance between Zipf plots. *Physica A: Statistical and Theoretical Physics* **216**, 148–150.

[14] Griffiths, T. L. (2011) Rethinking language: how probabilities shape the words we use. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 3825–6.

[15] Reilly, J & Kean, J. (2011) Information content and word frequency in natural language: word length matters. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E108; author reply E109.

[16] Piantadosi, S. T, Tily, H, & Gibson, E. (2011) Reply to Reilly and Kean: Clarifications on word length and information content. *Proceedings of the National Academy of Sciences* **108**, E109–E109.

[17] Reilly, J & Kean, J. (2007) Formal distinctiveness of high- and low-imageability nouns: analyses and theoretical implications. *Cognitive science* **31**, 157–68.

[18] Lazer, D, Pentland, A, Adamic, L, Aral, S, Barabasi, A.-L, Brewer, D, Christakis, N, Contractor, N, Fowler, J, Gutmann, M, Jebara, T, King, G, Macy, M, Roy, D, & Van Alstyne, M. (2009) Social science. Computational social science. *Science (New York, N.Y.)* **323**, 721–723.

[19] Brants, T & Franz, A. (2009) *Web 1T 5-gram, 10 European languages version 1.*

[20] Dodds, P. S & Danforth, C. M. (2009) Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies* **11**, 441–456.

[21] Russell, J. A. (1980) A circumplex model of affect. *Journal of personality and social psychology* **39**, 1161–1178.

[22] Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory*, Wiley Series in Telecommunications. (John Wiley & Sons, Inc., New York, USA) Vol. 6.

[23] Rime, B. (2009) Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review. *Emotion Review* **1**, 60–85.

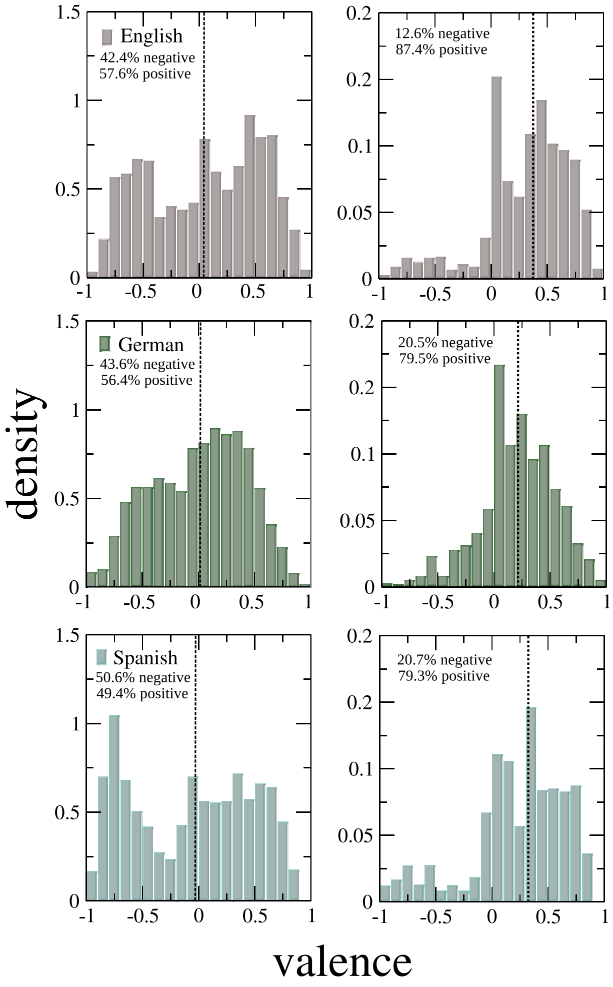

Figure 1: (left panel) Distributions of reported valence values for words in English (top panel, lexicon: [4], 1034 entries), German (middle panel, lexicon: [5], 2902 entries), and Spanish (bottom panel, lexicon: [6], 1034 entries), normalised by the size of the lexica. Average valence (median) 0.048 (0.095) for English, 0.021 (0.067) for German, and -0.065 (-0.006) for Spanish. (right panel) Normalised distributions of reported valence values weighted by the frequency of word usage, obtained from the same lexica. Average valence (median) 0.314 (0.375) for English, 0.200 (0.216) for German, and 0.238 (0.325) for Spanish. The dashed lines indicate the median. Inset numbers: ratio of positive and negative areas in the corresponding distributions.

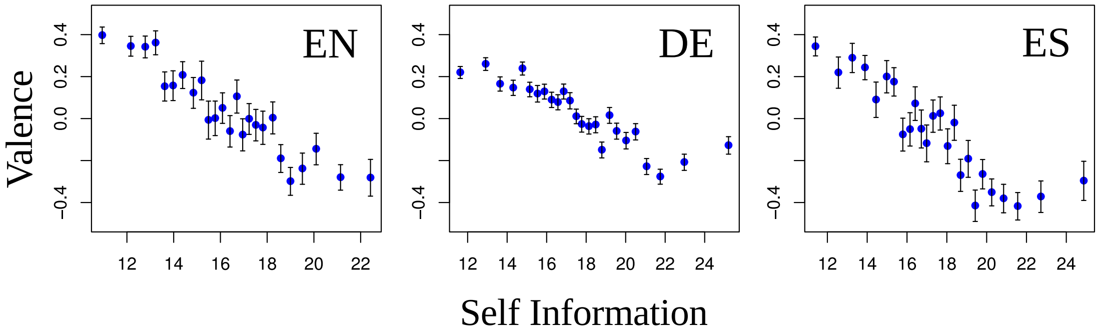

Figure 2: Relation between word self-information and valence for English (left), German (centre) and Spanish (right). Average valence is shown for bins that contain 5% of the data, with error bars showing the standard error. For all the three languages, valence clearly decreases with the self-information of the word, i.e. positive words carry less information than negative words.

# 5 Appendix

## 5.1 Correlations with self-information

We performed a detailed correlation analysis to understand the relations between valence ($v$), length ($l$) and self-information ($I$) for the three lexica. First, Pearson's correlation coefficient between word valence and self-information, $\rho(v, I)$, is significant and negative for the three languages (Table 1). This means that the higher the valence of a word, the lower its self-information. Second, we calculated Pearson's and Spearman's correlation coefficients between the absolute value of the valence and the self-information of a word, $\rho(abs(v), I)$. We found both correlation coefficients to be around 0.1 for German and Spanish, while they are not significant for English. The dependence between valence and self-information is destroyed if we ignore the sign of the valence.

| | English | German | Spanish |
|---|---|---|---|
| $\rho(v, I)$ | -0.368 ** | -0.325 ** | -0.402 ** |
| $\rho(abs(v), I)$ | 0.032 * | 0.109 ** | 0.135 ** |
| $\rho(l, I)$ | 0.378 ** | 0.143 ** | 0.361 ** |
| $\rho(v, l)$ | -0.044 * | -0.071 ** | -0.112 ** |
| $\rho(v, I\|l)$ | -0.379 ** | -0.319 ** | -0.399 ** |
| $\rho(l, I\|v)$ | 0.389 ** | 0.126 ** | 0.357 ** |

Significance levels: $^{**}p < 0.001$, $^{*}p < 0.3$.

Table 1: Correlation coefficients of the valence ($v$), absolute value of the valence ($abs(v)$), and word length ($l$) versus self-information ($I$). Partial correlations are calculated for both variables ($\rho(v, I|l)$,$\rho(l, I|v)$), and correlation between valence and length ($\rho(v, l)$).

For the three lexica, the correlation coefficient between word length and self-information ($\rho(l, I)$) is positive, showing that word length increases with self-information. These values of $\rho(l, I)$ are consistent with previous results [1, 2]. Pearson's and Spearman's correlation coefficients between valence and length $\rho(v, l)$ are very low or not significant. This supports the hypothesis that emotions are an additional component of the information transmitted.

In order to test the combined influence of valence and length to self-information, we calculated the partial correlation coefficients $\rho(v, I|l)$ and $\rho(l, I|v)$. The results are shown in Table 1, and are within the 95% confidence intervals of the original correlation coefficients $\rho(v, I)$ and $\rho(l, I)$.

## 5.2 Correlations with information content from $N$-grams and other word frequency datasets

In addition to self-information, we also calculated information content from entropy estimations based on the $N$-gram frequencies of the Google dataset. The original calculations were provided in Piantadosi et al. [2]. The correlation coefficient of valence and information content is similar for estimations using 2-grams and 3-grams. For the case of 4-grams they are comparable for English and Spanish, but not for German (Table 2). Each correlation coefficient becomes smaller for larger sizes of the context. This means that larger groups of words carry their own emotional information as a whole. The individual valence of the words might combine in a nonlinear way to create the whole sentence and text emotionality, for which a much larger survey study would be necessary, if even feasible.

| | English | German | Spanish |
|---|---|---|---|
| $\rho(v, I_2)$ | -0.332 ** | -0.301 ** | -0.359 ** |
| $\rho(v, I_3)$ | -0.313 ** | -0.201 ** | -0.359 ** |
| $\rho(v, I_4)$ | -0.254 ** | -0.049 * | -0.162 ** |
| $\rho(v, I')$ | -0.294 ** | -0.222 ** | -0.311 ** |

Significance levels: $^{*}p < 0.01$, $^{**}p < 0.001$.

Table 2: Correlation coefficients of the valence ($v$) and information content measured for 2-grams $I_2$, 3-grams $I_3$, and 4-grams $I_4$, and with self-information $I'$ measured from the frequencies reported in [4–6].

We provide additional support for our conclusions by estimating self-information from word frequency datasets independent of the Google $N$-grams we used. These frequency estimations, distributed with each lexicon [4–6], are based on texts of written communication, i.e. books. The correlation coefficients between word valence and these alternative estimations of self-information, $\rho(v, I')$, are consistent with our findings, as shown in Table 2. However, they are weaker and the frequencies are calculated over less tokens, but they provide robustness to our results on the self-information of word emotions.