Garcia et al. *EPJ Data Science* 2012, 1:3
http://www.epjdatascience.com/content/1/1/3

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR Open Access

Positive words carry less information than negative words

David Garcia*, Antonios Garas and Frank Schweitzer

*Correspondence: dgarcia@ethz.ch
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 analyzed 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.

PACS Codes: 89.65.-s; 89.70.Cf; 89.90.+n

1 Introduction

One would argue that human languages, in order to facilitate social relations, should be biased towards positive emotions. This question becomes particularly relevant for sentiment classification, as many tools assume as null hypothesis that human expression has neutral emotional content [4, 5], or reweight positive and negative emotions [6] without a quantification of the positive bias of emotional expression. We have tested and measured this bias in the context of online written communication by analyzing three established lexica of affective word usage. These lexica cover three of the most used languages on the Internet, namely English [7], German [8], and Spanish [9]. The emotional content averaged over all the words in each of them is 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 associated with a positive emotion are more frequently used than those associated with a negative emotion.

Historically, the *frequency* of words was first analyzed by Zipf [1, 10] showing that frequency predicts the *length* of a word as result of a principle of least effort. Zipf's law highlighted fundamental principles of organization in human language [11], and called for an interdisciplinary approach to understand its origin [12–14] and its relation to word meaning [15]. Recently Piantadosi et al. [2] extended Zipf's approach by showing that, in order to have efficient communication, word length increases with information content. Fur-

ther discussions [15–17] highlighted the relevance of *meaning* as part of the communication process as, for example, more abstract ideas are expressed through longer words [18]. Our work focuses on one particular aspect of meaning, namely the *emotion* expressed in a word, and how this is related to word frequency and information content. This approach requires additional data beyond word length and 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 [19]. Sentiment analysis techniques allow to quantify the emotions expressed through posts and messages [5, 6]. Recent studies have provided statistical analyses [20–23] and modelling approaches [24, 25] of individual and collective emotions on the Internet.

An emotional bias in written expressions, however, would have a strong impact, as it shifts the balance between positive and negative expressions. Thus, for all researchers dealing with emotions in written text it would be of particular importance to know about such bias, how it can be quantified, and how it affects the baseline, or reference point, for expressed emotions. Our investigation is devoted to this problem by combining two analyses, (i) quantifying the emotional content of words in terms of valence, and (ii) quantifying the frequency of word usage in the whole indexable web [26]. We provide a study of the baseline of written emotional expression on the Internet in three languages that span more than 67.7% of the websites [27]: English (56.6%), German (6.5%), and Spanish (4.6%). These languages are used everyday by more than 805 million users, who create the majority of the content available on the Internet.

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 importance of emotional content in human communication which influences the information carried by words. While the rational process that optimizes communication determines word lengths by the information they carry, 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. Furthermore, we extend the analysis of information content by taking into account word context rather than just word frequency. This leads to the conclusion that positive words carry less information than negative ones. In other words, the informativeness of words highly depends on their emotional polarity.

We wish to emphasize that our work distinguishes itself both regarding its methodology and its findings from a recent article [28]. There, the authors claim a bias in the amount of positive versus negative words in English, while no relation between emotionality and frequency of use was found. A critical examination of the conditions of that study shows that the quantification of emotions was done in an uncontrolled setup through the Amazon Mechanical Turk. Participants were shown a scale similar to the ones used in previous works [7–9], as explained in [23]. Thanks to the popular usage of the Mechanical Turk, the authors evaluated more than 10,000 terms from the higher frequency range in four different corpora of English expression. However, the authors did not report any selection criterion for the participant reports, opposed to the methodology presented in [29] where up to 50% of the participants had to be discarded in some experiments.

Because of this lack of control in their experimental setup, the positive bias found in [28] could be easily explained as an acquiescent bias [30, 31], a result of the human ten-

dency to agree in absence of further knowledge or relevance. In particular, this bias has been repeatedly shown to exist in self assessments of emotions [32, 33], requiring careful response formats, scales, and analyses to control for it. Additionally, the wording used to quantify word emotions in [28] (*happiness*), could imply two further methodological biases: The first one is a possible social desirability bias [34], as participants tend to modify their answers towards more socially acceptable answers. The positive social perception of displaying happiness can influence the answers given by the participants of the study. Second, the choice of the word *happiness* implies a difference compared with the standard psychological term *valence* [35]. Valence is interpreted as a static property of the word while happiness is understood as a dynamic property of the surveyed person when exposed to the word. This kind of framing effects have been shown to have a very large influence in survey results. For example, a recent study [36] showed a large change in the answers by simply changing *voting* for *being a voter* in a voter turnout survey.

Hence, there is a strong sensitivity to such influences which are not controlled for in [28]. Because of all these limitations, in our analysis we chose to use the current standard lexica of word valence. These lexica, albeit limited to 1,000 to 3,000 words, were produced in three controlled, independent setups, and provide the most reliable estimation of word emotionality for our analysis. Our results on these lexica are consistent with recent works on the relation between emotion and word frequency [37, 38] for English in corpora of limited size.

2 Results

2.1 Frequency of emotional words

In detail, we have analyzed three lexica of affective word usage which contain 1,034 English words, 2,902 German words and 1,034 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 [39]. Each word in these lexica is assigned a set of values measuring different aspects of word emotionality. The three independent studies that generated the lexica for English [7], German [8], and Spanish [9] used the Self-Assessment Mannequin (SAM) method to ask participants about the different emotion values associated to each word in the lexicon. One of these values, a scalar variable v called valence, represents the degree of pleasure induced by the emotion associated with the word, and it is known to explain most of the variance in emotional meaning [35]. In this article, we use v to quantify word emotionality.

In each lexicon, words were chosen such that they evenly span the full range of valence.[a] 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 Figure 2, 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, 10]. For our frequency estimations we have used Google's N-gram dataset [26] which, with 10^{12} tokens, is one of the largest datasets available about real human text expressions on the Internet. For our analysis, we have studied the frequency of the words which have an affective classification in the respective lexicon in either English, German, or Spanish. Figure 1 shows emotion word clouds for the three languages, where each word appears with a size proportional to its frequency. The color of a word is chosen according to its valence, ranging from red for

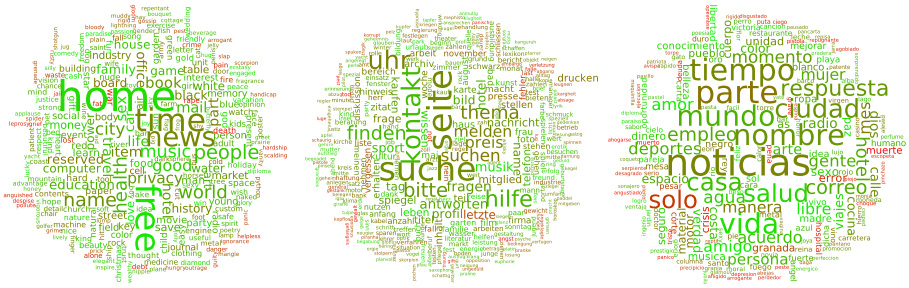

Figure 1 Emotion word clouds with frequencies calculated from Google's crawl. In each word cloud for English (left), German (middle), and Spanish (right), the size of a word is proportional to its frequency of appearance in the trillion-token Google *N*-gram dataset [26]. Word colors are chosen from red (negative) to green (positive) in the valence range from psychology studies [7–9]. For the three languages, positive words predominate on the Internet.

$\nu = -1$ to green for $\nu = +1$. It is clear that green dominates over red in the three cases, as positive emotions predominate on the Internet. Some outliers, like "home", have a special higher frequency of appearance in websites, but as we show later, our results are consistent with frequencies measured from traditional written texts like books.

In a general setup, 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 Figure 2. Wilcoxon tests show that the means of these distributions are indeed different, with an estimated difference in a 95% confidence interval of 0.257 ± 0.032 for English, 0.167 ± 0.017 for German, and 0.287 ± 0.035 for Spanish. Hence, with respect to usage we find evidence that the language used on the Internet is 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).

2.2 Relation between information and valence

Our analysis suggests that there is a definite relation between word valence and frequency of use. 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 can approximate its value given the frequencies of words and word sequences. First we discuss the relation between word valence and information content estimated from the simple word occurrences, namely self-information. Then we explain how this extends when the information is measured taking into account the different contexts in which a word can appear. The self-information of a word, $I(w)$ [40] is an estimation of the information content from its probability of appearance, $P(w)$, as

$$I(w) = -\log P(w) \tag{1}$$

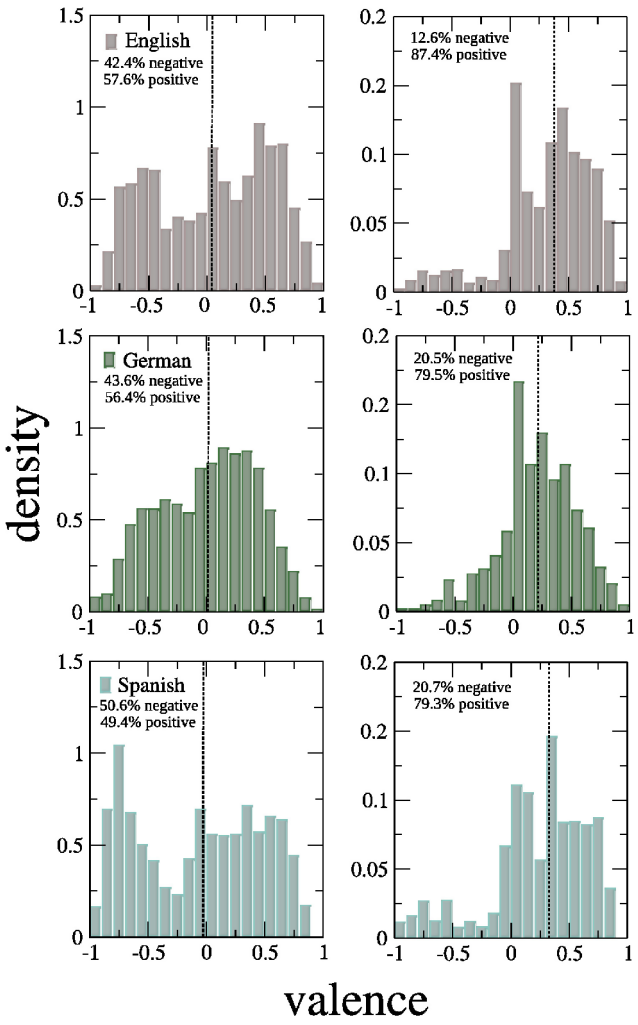

Figure 2 Distributions of word emotions weighted by the frequency of word usage. (left panel) Distributions of reported valence values for words in English (top panel, lexicon: [7], 1,034 entries), German (middle panel, lexicon: [8], 2,902 entries), and Spanish (bottom panel, lexicon: [9], 1,034 entries), normalized 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) Normalized 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.

Frequency-based information content metrics like self-information are commonly used in computational linguistics to systematically analyze communication processes. Information content is a better predictor for word length than word frequency [2, 41], and the relation between information content and meaning, including emotional content, is claimed to be crucial for the way humans communicate [15–17]. We use the self-information of a word as an estimation of information content for a context size of 1, to build up later on larger context sizes. This way, we frame our analysis inside the larger framework of N-gram information measures, aiming at an extensible approach that can be incorporated in the fields of computational linguistics and sentiment analysis.

For the three lexica, we calculated $I(w)$ of each word and linked it to its valence, $\nu(w)$. As defined in Equation 1, very common words provide less information than very unusual

Table 1 Correlations between word valence and information measurements.

	English	German	Spanish
$\rho(v,f)$	0.222^{**}	0.144^{**}	0.236^{**}
$\rho(v,I)$	-0.368^{**}	-0.325^{**}	-0.402^{**}
$\rho(v,I')$	-0.294^{**}	-0.222^{**}	-0.311^{**}
$\rho(v,I_2)$	-0.332^{**}	-0.301^{**}	-0.359^{**}
$\rho(v,I_3)$	-0.313^{**}	-0.201^{**}	-0.340^{**}
$\rho(v,I_4)$	-0.254^{**}	-0.049^{*}	-0.162^{**}

Correlation coefficients of the valence (v), frequency f, self-information I, 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 [42–44]. Significance levels: $^{*}p < 0.01$, $^{**}p < 0.001$.

ones, but this nonlinear mapping between frequency and self-information makes the latter more closely related to word valence than the former. The first two lines of Table 1 show the Pearson's correlation coefficient of word valence and frequency $\rho(v,f)$, followed by the correlation coefficient between word valence and self-information, $\rho(v,I)$. For all three languages, the absolute value of the correlation coefficient with I is larger than with f, showing that self-information provides more knowledge about word valence than plain frequency of use.

The right column of Figure 3 shows in detail the relation between v and I. 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. As mentioned before the Pearson's correlation coefficient between word valence and self-information, $\rho(v,I)$, is significant and negative for the three languages (Table 1). Our results outperform a recent finding [37] that, while focusing on individual text production, reported a weaker correlation (below 0.3) between the logarithm of word usage frequency and valence. This previous 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 [7] 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.

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, i.e. books, as reported in the original datasets for English [42], German [43], and Spanish [44]. Calculating the self-information from these and relating them to the valences given, we obtained similar, but slightly lower Pearson's correlation coefficients $\rho(v,I')$ (see Table 1). So, we conclude that our results are robust across different types of written communication, for the three languages analyzed.

It is not surprising to find a larger self-information for negative words, as their probability of appearance is generally lower. The amount of information carried by a word is also highly dependent on its context. Among other factors, the context is defined by the word neighborhood in the sentence. For example, the word "violent" contains less information in the sentence "dangerous murderers are violent" than in "fluffy bunnies are violent", as the probability of finding this particular word is larger when talking about murderers than about bunnies. For this reason we evaluate how the context of a word impacts its informativeness and valence. The intuition behind measuring information depending on the

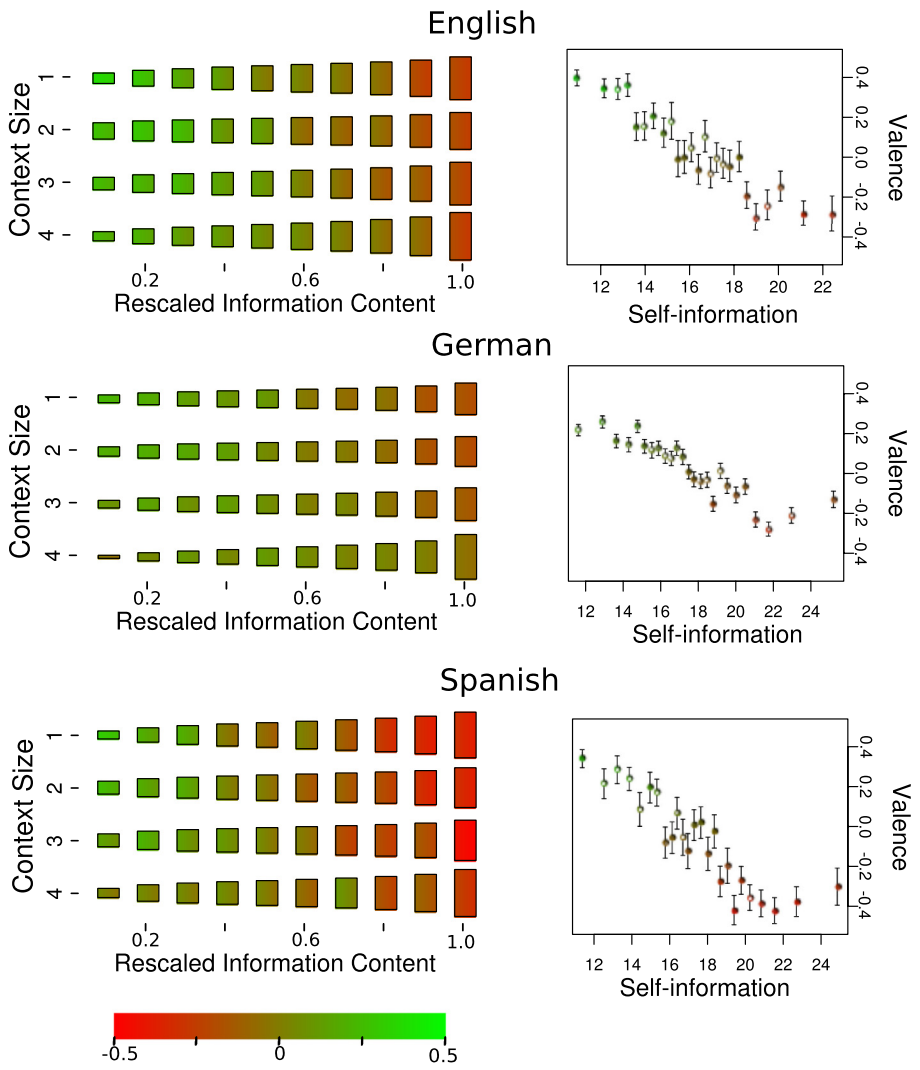

Figure 3 Relation between information measures and valence. Each graphic on the left column shows the relation between valence and information content measured up to a context of size four. Each bar represents a bin containing 10% of the words in the lexicon, with a size proportional to the average information content of the words in the bar. The color of each bar ranges from red to green, representing the average valence of the words in the corresponding bin. Each bar has a color gradient according to the standard error of the valence mean. Information content has been rescaled so it can be compared among context sizes. For all three languages and context sizes, negativity increases with information content. The second column shows the relation between word self-information and valence for English, German, and Spanish. 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.

context is that the information content of a word depends primarily on i) the amount of contexts it can appear and ii) the probability of appearance in each one of these contexts. Not only the most infrequent, but the most specific and unexpectable words are the ones that carry the most information. Given each context c_i where a word w appears, the information content is defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\log(\mathrm{P}(W=w|C=c_i)) \tag{2}$$

where N is the total frequency of the word in the corpus used for the estimation. These values were calculated as approximations of the information content given the words surrounding w up to size 4.

We analyzed how word valence is related to the information content up to context size 4 using the original calculations provided by Piantadosi et al. [2]. This estimation is based on the frequency of sequences of N words, called N-grams, from the Google dataset [26] for $N \in \{2, 3, 4\}$. This dataset contains frequencies for single words and N-grams, 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. For each size of the context N, we have a different estimation of the information carried by the studied words, with self-information representing the estimation from a context of size 1.

The left column of Figure 3 shows how valence decreases with the estimation of the information content for each context size. Each bar represents the same amount of words within a language and has an area proportional to the rescaled average information content carried by these words. The color of each bar represents the average valence of the binned words. The decrease of average valence with information content is similar for estimations using 2-grams and 3-grams. For the case of 4-grams it also decreases for English and Spanish, but this trend is not so clear for German. These trends are properly quantified by Pearson's correlation coefficients between valence and information content for each context size (Table 1). Each correlation coefficient becomes smaller for larger sizes of the context, as the information content estimation includes a larger context but becomes less accurate.

2.3 Additional analysis of valence, length and self-information

In order to provide additional support for our results, we tested different hypotheses impacting the relation between word usage and valence. First, we calculated Pearson's and Spearman's correlation coefficients between the absolute value of the valence and the self-information of a word, $\rho(\text{abs}(v), I)$ (see Table 2). 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 disappears if we ignore the sign of the valence, 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 by the word.

Subsequently, we found that 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

Table 2 Additional correlations between valence, self-information and length.

	English	German	Spanish
$\rho(\text{abs}(v), I)$	0.032°	0.109^{***}	0.135^{***}
$\rho(l, I)$	0.378^{***}	0.143^{***}	0.361^{***}
$\rho(v, I)$	-0.044°	-0.071^{***}	-0.112^{***}
$\rho(v, l \mid I)$	-0.379^{***}	-0.319^{***}	-0.399^{***}
$\rho(l, l \mid v)$	0.389^{***}	0.126^{***}	0.357^{***}

Correlation coefficients of the valence (v), absolute value of the valence ($\text{abs}(v)$), and word length (l) versus self-information (I). Partial correlations are calculated for both variables ($\rho(v, l \mid I)$,$\rho(l, I \mid v)$), and correlation between valence and length ($\rho(v, l)$). Significance levels: $^{\circ}p < 0.3$, $^{*}p < 0.1$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 3 Partial correlation coefficients between valence and information content.

	English	German	Spanish
$\rho(v, l_2 \mid l)$	-0.034°	-0.100^{***}	-0.058^{*}
$\rho(v, l_3 \mid l)$	-0.101^{**}	-0.070^{***}	-0.149^{***}
$\rho(v, l_4 \mid l)$	-0.134^{***}	-0.020^{*}	-0.084^{**}

Correlation coefficients of the valence (v) and information content measured on different context sizes (l_2, l_3, l_4) controlling for self-information (l). Significance levels: $^{\circ}p < 0.3$, $^{*}p < 0.1$, $^{**}p < 0.01$, $^{***}p < 0.001$.

correlation coefficients between valence and length $\rho(v, l)$ are very low or not significant. 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 2, and are within the 95% confidence intervals of the original correlation coefficients $\rho(v, I)$ and $\rho(l, I)$. This provides support for the existence of an additional dimension in the communication process closely related to emotional content rather than communication efficiency. This is consistent with the known result that word lengths adapt to information content [2], and we discover the independent semantic feature of valence. Valence is also related to information content but not to the symbolic representation of the word through its length.

Finally, we explore the sole influence of context by controlling for word frequency. In Table 3 we show the partial correlation coefficients of valence with information content for context sizes between 2 and 4, controlling for self-information. We find that most of the correlations keep significant and of negative sign, with the exception of I_2 for English. The weaker correlation for context sizes of 2 is probably related to two word constructions such as negations, articles before nouns, or epithets. These high-frequent, low-information constructions lead to the conclusion that I_2 does not explain more about the valence than self-information in English, as short range word interactions change the valence of the whole particle. This finding supports the assumption of many lexicon-based unsupervised sentiment analysis tools, which consider valence modifiers for two-word constructions [5, 6]. On the other hand, the significant partial correlation coefficients with I_3 and I_4 suggest that word information content combines at distances longer than 2, as longer word constructions convey more contextual information than 2-grams. Knowing the possible contexts of a word up to distance 4 provides further information about word valence than sole self-information.

3 Discussion

Our analysis provides strong evidence that words with a positive emotional content are used more often. 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 information, measured in terms of self-information, compared to positive words. This relation remains valid even when considering the context composed of sequences of 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 [38]. The question of how syntax influences emotional expression is beyond the scope of the present work. Note that the lexica we use are composed mainly of nouns, verbs and adjectives, due to their emotional relevance. Function words such as "a" or "the" are not considered to have any emotional content and therefore were excluded from the original studies. In isolation, these function words do not contain explicit valence content, but their presence in text can modify the meaning of neighboring words and thus modify the emotional content of a sentence as a whole. Our analysis on partial correlations show that there is a correlation between the structure of a sentence and emotional content beyond the simple appearance of individual words. This result suggests the important role of syntax in the process of emotional communication. Future studies can extend our analysis by incorporating valence scores for word sequences, exploring how syntactical rules represent the link between context and emotional content.

The findings reported in this paper suggest that the process of communication between humans, which is known to optimize information transfer [2], also creates a bias towards positive emotional content. A possible explanation is the 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. Thus, it makes much sense that human languages on average have strong bias towards positive emotions, as we have shown (see Figure 2). 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 consistent with emotion research in social psychology. According to [45], 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. Our analysis provides insights on the asymmetry of evaluative processes, as frequent positive expression is consistent with the concept of *positivity offset* introduced in [46] and recently reviewed in [47]. In addition, Miller's *negativity bias* (stronger influence of proximal negative stimuli) found in experiments provides an explanation for the higher information content of negative expression. When writing, people could have a tendency to avoid certain negative topics and bring up positive ones just because it feels better to talk about nice things. That would lower the frequency of negative words and lower the amount of information carried by positive expression, as negative expression would be necessary to transmit information about urgent threats and dangerous events.

Eventually, we emphasize that the positive emotional "charge" of human communication has a further impact on the quantitative analysis of communication on the Internet, for example in chatrooms, forums, 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. In addition, this relation between

information content and word valence might be useful to detect anomalies in human emotional expression. Fake texts supposed to be written by humans could be detected, as they might not be able to reproduce this spontaneous balance between information content and positive expression.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
All authors designed and performed research, analyzed data, and wrote the article.

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

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

Recieved: 27 January 2012 Accepted: 18 May 2012 Published: 18 May 2012

References

1. Zipf GK (1935) The psycho-biology of language. Houghton Mifflin, Oxford
2. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. Proc Natl Acad Sci USA 108(9):3526
3. Boucher J, Osgood C (1969) The Pollyanna hypothesis. J Verbal Learn Verbal Behav 8(1):1-8
4. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(12):1-135
5. Thelwall M, Buckley K, Paltoglou G (2011) Sentiment in Twitter events. J Am Soc Inf Sci Technol 62(2):406-418
6. Taboada M, Brooke J, Voll K (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267-307
7. Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida
8. Võ ML-H, Conrad M, Kuchinke L, Urton K, Hofmann MJ, Jacobs AM (2009) The Berlin affective word list reloaded (BAWL-R). Behav Res Methods 41(2):534-538
9. Redondo J, Fraga I, Padrón I, Comesaña M (2007) The Spanish adaptation of ANEW (affective norms for English words). Behav Res Methods 39(3):600-605
10. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, New York
11. Ferrer i Cancho R, Sole RV (2003) Least effort and the origins of scaling in human language. Proc Natl Acad Sci USA 100(3):788-791
12. Hauser MD, Chomsky N, Fitch WT (2002) The faculty of language: what is it, who has it, and how did it evolve? Science 298(5598):1569-1579
13. Kosmidis K, Kalampokis A, Argyrakis P (2006) Statistical mechanical approach to human language. Physica A 366:495-502
14. Havlin S (1995) The distance between Zipf plots. Physica A 216(1-2):148-150
15. Piantadosi ST, Tily H, Gibson E (2011) Reply to Reilly and Kean: clarifications on word length and information content. Proc Natl Acad Sci USA 108(20):E109-E109
16. Griffiths TL (2011) Rethinking language: how probabilities shape the words we use. Proc Natl Acad Sci USA 108(10):3825-3826
17. Reilly J, Kean J (2011) Information content and word frequency in natural language: word length matters. Proc Natl Acad Sci USA 108(20):E108; author reply E109
18. Reilly J, Kean J (2007) Formal distinctiveness of high- and low-imageability nouns: analyses and theoretical implications. Cogn Sci 31(1):157-168
19. 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 323(5915):721-723
20. Bollen J, Mao H, Zeng X-j (2011) Twitter mood predicts the stock market. Journal of Computational Science 2:1-8
21. Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051):1878-1881
22. Chmiel A, Sienkiewicz J, Thelwall M, Paltoglou G, Buckley K, Kappas A, a Hoyst J (2011) Collective emotions online and their influence on community life. PLoS ONE 6(7):e22207
23. Dodds PS, Harris KD, Kloumann IM, a Bliss C, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PLoS ONE 6(12):e26752
24. Schweitzer F, Garcia D (2010) An agent-based model of collective emotions in online communities. Eur Phys J B 77(4):533-545
25. Garcia D, Schweitzer F (2011) Emotions in product reviews – empirics and models. In: Bilof R (ed) Proceedings of 2011 IEEE international conference on privacy, security, risk, and trust, and IEEE international conference on social computing, PASSAT/SocialCom. IEEE Computer Society, Boston, pp 483-488
26. Brants T, Franz A (2009) Web 1T 5-gram, 10 European languages version 1. Linguistic Data Consortium. Philadelphia
27. Global internet usage. en.wikipedia.org/wiki/Global_Internet_usage
28. Kloumann IM, Danforth CM, Harris KD, a Bliss C, Dodds PS (2012) Positivity of the English language. PLoS ONE 7(1):e29484

29. Bohannon J (2011) Social science for pennies. Science, 334(October):2011
30. Knowles ES, Nathan KT (1997) Acquiescent responding in self-reports: cognitive style or social concern? J Res Pers 301(31):293-301
31. Bentler PM (1969) Semantic space is (approximately) bipolar. J Psychol 71(1):33-40
32. Russell JA (1979) Affective space is bipolar. J Pers Soc Psychol 37(3):345-356
33. Yik M, a Russell J, Steiger JH (2011) A 12-point circumplex structure of core affect. Emotion 11(4):705-731
34. Robinson JP, Shaver PR, Wrightsman LS (1991) Measures of personality and social psychological attitudes. Elsevier, Amsterdam
35. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161-1178
36. Bryan CJ, Walton GM, Rogers T, Dweck CS (2011) Motivating voter turnout by invoking the self. Proc Natl Acad Sci USA 108(31):12653-12656
37. Augustine AA, Mehl MR, Larsen RJ (2011) A positivity bias in written and spoken English and its moderation by personality and gender. Social Psychological and Personality Science 2(5):508-515
38. Rozin P, Berman L, Royzman E (2010) Biases in use of positive and negative words across twenty natural languages. Cogn Emot 24(3):536-548
39. Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. Journal of Happiness Studies 11(4):441-456
40. Cover TM, Thomas JA (1991) Elements of Information theory. Wiley series in telecommunications, vol 6. Wiley, New York
41. Jaeger TF (2010) Redundancy reduction: speakers manage syntactic information density. Cogn Psychol 61(1):23-62
42. Kucera H, Francis WN (1967) Computational analysis of presentday American English. Brown University Press, Providence
43. Baayen RH, Piepenbrock R, van Rijn H (1993) The CELEX lexical database [CD-ROM]. Technical report, University of Pennsylvania, Linguistic Data Consortium, Philadelphia
44. Sebastián N, Martí MA, Carreiras MF, Cuetos F (2000) LEXESP: Léxico informatizado del español. Ediciones de la Universitat de Barcelona, Barcelona
45. Rime B (2009) Emotion elicits the social sharing of emotion: theory and empirical review. Emotion Review 1(1):60-85
46. Miller NE (1961) Some recent studies of conflict behavior and drugs. Am Psychol 16(1):12-24
47. Norman GJ, Norris CJ, Gollan J, a Ito T, Hawkley LC, Larsen JT, Cacioppo JT, Berntson GG (2011) Current emotion research in psychophysiology: the neurobiology of evaluative bivalence. Emotion Review 3(3):349-359

doi:10.1140/epjds3
Cite this article as: Garcia et al.: **Positive words carry less information than negative words.** *EPJ Data Science* 2012 **1**:3.