Robustness of emotion extraction from 20th century English books

Alberto Acerbi
Department of Archaeology and Anthropology
University of Bristol
Bristol, United Kingdom
Email: alberto.acerbi@gmail.com

Vasileios Lampos
Department of Computer Science
University of Sheffield
Sheffield, United Kingdom
Email: bill@lampos.net

R. Alexander Bentley
Department of Archaeology and Anthropology
University of Bristol
Bristol, United Kingdom
Email: r.a.bentley@bristol.ac.uk

***Abstract*—In this paper, we test the robustness of emotion extraction from English language books published in the 20th century. Our analysis is performed on a sample of the 8 million digitized books available in the Google Books Ngram corpus by applying three independent emotion detection tools: WordNet Affect, Linguistic Inquiry and Word Count, and a recently proposed 'Hedonometer' method. We also assess the statistical robustness of the extracted patterns as well as their outputs on specific parts of speech. The analysis confirms three main results: the existence of recognizable periods of positive and negative 'literary affect' from 1900 to 2000, a general decrease in the usage of emotion-related words in printed books that lasts at least until the 1980s, and, finally, a divergence between American and British books, with the former using more emotion-related words from the 1960s.**

I. Introduction

The study of cultural dynamics has recently been transformed by the availability, and by the relative ease of storage and analysis, of massive amounts of data. Novel forms of human-generated input (*e.g.*, Facebook, Twitter, blogs) are produced daily, forming an interesting information source for studies on social and cultural behavior. At the same time, an increasing amount of 'traditional' data, such as books and newspaper articles, is digitized and made available for quantitative analysis.

One criticism targeting the use of 'Big Data' in social and human sciences is that the vast majority of works focuses on a short-time scale. However, 'Long Data' [1] with a temporal span of years or even centuries (as opposed to days or months for Social Media studies) are available as well and could be potentially used to answer different questions. For example, a new field of research dubbed 'Culturomics' [2] proposes to use quantitative data, in particular word frequencies in millions of digitized books, to help understand aspects of cultural dynamics at longer time scales.

Building on those ideas, recent studies tried to analyse the use of emotion-related words on long-time scale. An analysis of song lyrics, for example, showed a downward trend in their 'happiness' from the 1960s to the mid 1990s [3]. DeWall *et al.* [4], similarly, studied word usage in song lyrics from 1980 to 2007, and found that the use of 'angry' and 'antisocial' lyrics increased over this period. In a follow-up study, Twenge *et al.* [5] concluded that individualistic words increased in American books between 1960 and 2008, whereas communal words did not. In another example, it was shown that fairy tales have a much wider range of emotion word densities than novels, and thus, are generally more 'emotional' [6].

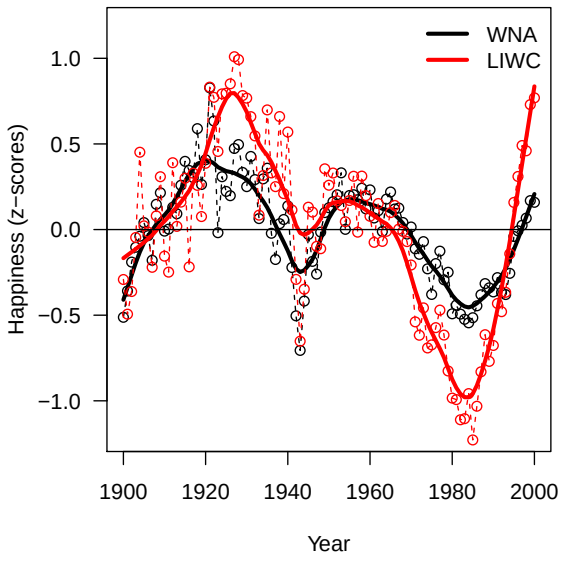

Fig. 1: 'Happiness' z-scores for years 1900 to 2000 (circles show actual times series, the lines show the smoothed time series) under WNA and LIWC. Values above zero indicate generally 'happy' periods, and values below the zero indicate generally 'sad' periods.

In our recent work [7], we used the Google Books Ngram corpus [2] to extract emotional trends from 20th century books written in English language. Our analysis yielded three main results: a) the existence of recognizable periods of positive and negative affect (see an example in Figure 1), b) a general decrease in the use of emotion-related words throughout the century, and c) a divergence between American and British books, with the former being comparatively more emotional from the 1960s onwards.

In this paper, we report on a series of analyses performed in order to check the level of robustness in the emotional scoring

of books. Primarily, we have updated our findings to include the new version of the Google Books Ngram corpus, available from July 2012, which contains data from more than 8 million digitized books[1] [8].

An obvious concern regards the suitability of the tools applied for the emotion extraction from text. In [7], we applied WordNet Affect (WNA), a variant of WordNet which consists of emotion-oriented words extracted by selecting and labeling synsets representing affective concepts [9]–[11]. Previously, WNA has been successfully applied for the task of extracting collective mood patterns from human-generated content posted on Social Media, such as the microblogging platform of Twitter [12]–[14]. Here, we compare WNA results with the outputs obtained from two alternative mood extraction tools, namely the Linguistic Inquiry and Word Count (LIWC) ([15], [16]), and a recently developed method for extracting the degree of 'Happiness' from content posted on Twitter, dubbed 'Hedonometer' (HED) [17]. LIWC has also been incorporated in recent research developments, such as models for predicting an election outcome [18] or for estimating circadian patterns of affect [19] based on Social Media content.

Furthermore, we assess the statistical robustness of the extracted emotion patterns by estimating confidence intervals (CIs) for the central results. As it is well known, the distribution of word frequencies in language follows 'Zipf's law', where the frequency of a word is inversely proportional to its frequency rank [20]–[22]. For our results, this means that high-frequency terms might determine on their own the trends for specific emotions, obscuring the role of the numerous low-frequency terms.

Lastly, an aspect overlooked in our previous analysis was the role of Part-Of-Speech (POS) information. It has been proposed that certain lexical categories, as adjectives and adverbs, are particularly good indicators of emotional content [23]. The new version of the Google Books Ngram corpus provides POS tags, that were absent from the previous version [8], hence, we now can compare our original results with trends obtained by considering only terms tagged as adjectives or adverbs.

II. Data

We use data provided in the second version of the Google Books Ngram corpus.[2] The corpus is a digitization of $8,116,746$ volumes, which represent approximatively a 6% of all books ever published [8]. Books from a variety of languages are included, but, for our analysis, we make use of English language books divided in three sub-corpora and we limit our queries to volumes published between 1900 and 2000 (both included). The three sub-corpora considered are: English (all books, for a total, in the period considered, of $2,980,271$ volumes), American English (English language books published in United States; $2,073,315$ volumes), and, finally, British English (English language books published in Great Britain; $796,363$ volumes).

[1]The version used in our initial analysis [7] contained approx. 5 million books.

[2]Google Books Ngram data sets, http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

The corpora give 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, and so on) and an n-gram is a sequence of n 1-grams. For our analysis we use only the frequencies of 1-grams since the emotion extraction lexicons are also based around single terms. Additionally, the corpora provide syntactic annotations, by tagging words with their POS [8].

III. Methods

In this section, we describe the three emotion detection tools utilised (WNA, LIWC, HED) as well as the filtering procedure of the Google corpus, and the additional analysis performed to compute CIs and to account for the various POS.

We focus on the three results mentioned in the Introduction. For the first one (*i.e.,* existence of recognizable periods of positive and negative affect), we compare the outputs of the three emotion detection tools. For the second (decrease of emotion-related words) and third results (American and British English comparison) we can only use WNA and LIWC, as HED does not provide an assessment of the general emotional content in the texts. The additional analyses (CIs and POS analysis) are performed on both WNA and LIWC for the first result, and on WNA only for the second and the third ones.

A. Emotion Detection Tools

In our previous work [7] we have used, to extract emotions from the textual content of the books, WNA, a taxonomy of affective terms [10]. For our analysis we have considered a Porter-stemmed [24] version of WNA, which includes six mood categories, each represented by a different number of terms: Anger ($N = 146$), Disgust ($N = 30$), Fear ($N = 92$), Joy ($N = 224$), Sadness ($N = 115$), and Surprise ($N = 41$).

Here we validate our results using a different emotion detection tool (LIWC), which, unlike WNA, is a taxonomy of affective terms that has been evaluated by human judges [15]. LIWC already includes stems of words together with complete (non-stemmed) words in all of its vocabularies. We consider the LIWC categories of General Affect ($N = 917$), Anger ($N = 184$), Anxiety ($N = 91$), Negative Emotions ($N = 499$), Positive Emotions ($N = 408$), and Sadness ($N = 101$).

Furthermore, to quantify happiness in books we also apply the 'Hedonometer' method (HED) [17]. This metric, applied also on Twitter data to extract various patterns of collective mood, is based on a set of $3,686$ weighted words which were previously evaluated for their degree of happiness using Amazon's Mechanical Turk. Differently from WNA and LIWC, HED does not consider explicitly terms with emotional content, but words chosen by frequency of usage and then evaluated for their degree of 'happiness', so that one can find terms such as "food" or "Christmas" (happy) and "funeral" or "terrorism" (sad). Another difference is that, in HED, terms are weighted, so that, if two terms have the same frequency in a text, they will anyway contribute differently to the general 'Happiness' score of that text [17].

B. Vector Space Model

Similar to [25], we represent the word frequency data by vectors, which, in our case, correspond to emotion words in a given data set for a particular year.

We compute 1-gram frequencies from 1900 to 2000 (a total of 101 years) from the Google Books Ngram corpus on three data sets: a) all books in English language, b) all books written in American English (published in the USA), and c) all books written in British English (published in the UK).

Google Books corpus provides the total count of 1-grams for each year. Since the number of books present in the corpus varies considerably through the years (books for 2000 are about 10 times more than for the beginning of the century), we obtain frequencies by normalizing the yearly count 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 ([7], [26]).

For a year Y, given the count C_{the} of the word "the" in the corpus as well as the counts $\{c_i, ..., c_n\}$ of the n terms (case-insensitive) representing a mood type, we compute a mood score ($\mathcal{M}_Y$) as follows:

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

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 convert them to their z-score equivalent ($\mathcal{M}z_Y$), using:

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}}, \tag{2}$$

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

In the case of HED, we extract from the Google Books corpus the frequencies of $3,686$ words taken from a list of $10,222$ words provided in [17]. These terms represent the words that were evaluated as particularly 'happy' or particularly 'unhappy', and exclude 'neutral', high-frequency words (stop-words). Their yearly (case-insensitive) counts are normalised using the total yearly sum (in this case, in order to obtain results comparable with [17], we do not use the count of "the"), multiplied with the weights provided in [17], and finally summed.

To avoid any bias related to normalization in determining the 'absolute' trends of moods (for result b – decrease in the use of emotion-related words), we repeated the procedure described in [7], comparing the z-scores of moods time series with a z-score derived by a random sample of $10,000$ terms extracted from the Part of Speech database.[3] For the WNA experiments this random sample of terms is stemmed, whereas for LIWC we use a mix of stemmed and non-stemmed random terms, to replicate the composition of the lists.

[3]Part of Speech database, http://wordlist.sourceforge.net/pos-readme.

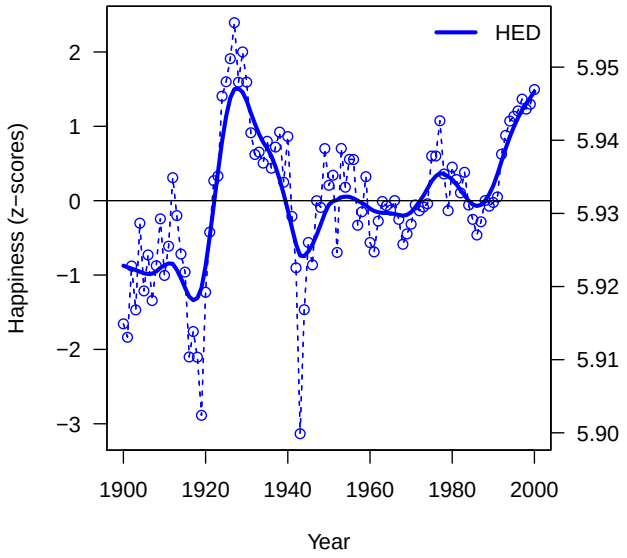

Fig. 2: 'Happiness' z-scores for years 1900 to 2000 using HED (circles show actual times series, the line shows the smoothed time series). Values above zero indicate generally 'happy' periods, and values below the zero indicate generally 'sad' periods. The right y-axis reports the absolute values of the 'Hedonometer'.

C. Confidence Interval Estimation and Part-Of-Speech Analysis

To assess the robustness of our results, CIs are estimated for the main emotion patterns we have presented. For this purpose, we apply bootstrap sampling [27]. The space of mood words is sampled with replacement $10,000$ times, *i.e.*, in each bootstrap, we are using a random subset of the emotion terms to compute the emotion time series (or a difference between two time series, where applicable). We average across those results in a yearly fashion to retrieve a mean time series and use the 0.025 and 0.975 quantiles of the bootstrap samples to derive 95% CIs. The latter, consequently, may vary per year.

As a further quality control to our initial results, we repeat our main analysis tasks, this time taking in consideration only terms that were tagged in the Google corpus as adjectives (ADJ) or adverbs (ADV). The rational behind this action is based on the experimentally proven hypothesis that adverbs and adjectives are probably the best indicators of emotional content [23].

IV. Results

In the next paragraphs, we present all the results derived from our analysis. Our plots also contain a smoothed trend which is computed using Friedman's 'super smoother' [28] through R's function *supsmu*.[4] Comparisons between different time series are always indicated using Pearson's correlations with a sample size of $N = 101$ (*i.e.*, the years $\in [1900, 2000]$).

[4]The R Project for Statistical Computing, http://www.r-project.org/.

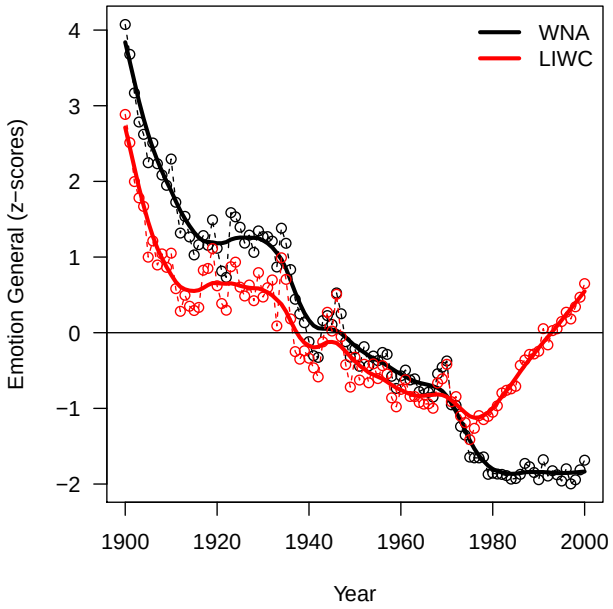

Fig. 3: WNA: difference between z-scores of an aggregation of all six emotions and of a random sample of stemmed words. LIWC: difference between z-scores of General Affect and of a random sample of stemmed and non-stemmed words. Circles show actual times series, the lines show the smoothed time series.

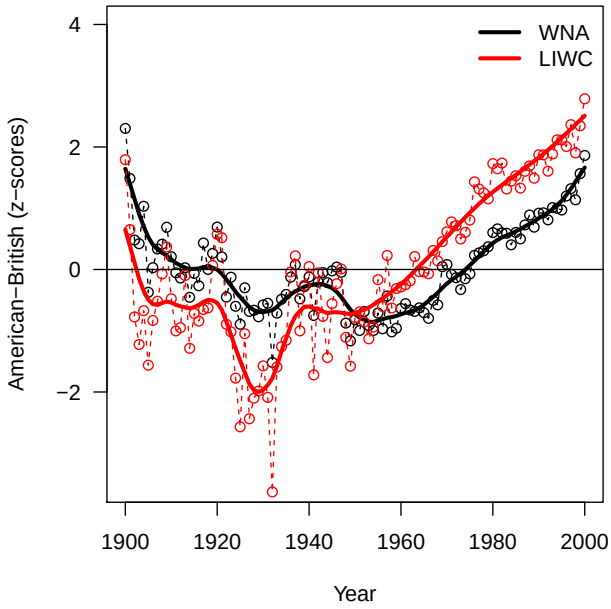

Fig. 4: Difference between z-scores for emotion terms in American English and British English (circles show actual times series, the lines show the smoothed time series). For WNA we aggregated all six emotions, whereas for LIWC we used the category of General Affect.

A. *Comparison of emotion detection tools*

We compute, as in [7], a general 'Happiness' index for the 20^{th} century books present in the Google Books Ngram corpus. For WNA, this index is obtained as the difference between the z-score of Joy and the z-score of Sadness. Analogously, for LIWC, the index is obtained as the difference between the z-scores of terms representing Positive Emotions and Sadness. The two indices (see Figure 1) are strongly correlated (Pearson's $r = 0.82$, $p < 0.0005$).

Figure 2 shows the same 'Happiness' index calculated trough HED. While some similarities are present (*e.g.*, the 'happiness' peak at the end of 1920s and the 'sadness' peak corresponding to the Second World War), the index does not correlate with the respective WNA time series (Pearson's $r = 0.18$, $p = 0.07$). A straightforward observation for this index is that, in contrast with the previous results, manages to track a negative period corresponding to the First World War (compare Figures 1 and 2).

The second trend we analyse is the general usage of emotion-related terms in the whole century. Figure 3 shows that both WNA and LIWC identify a steady decrease in the use of affective words trough the century that stops (WNA) or changes direction (LIWC) in the last two decades. The time series produced by WNA and LIWC are highly correlated (Pearson's $r = 0.81$, $p < 0.0005$), however they deviate clearly from the 1980s onwards. The main explanation for this resides on the fact that LIWC contains a multitude of modern and possibly less formal terms that express emotion. For example, the top-10 LIWC entries in terms of relative frequency increase in that period are: *geek**, *soulmate**, *ROFL*, *laidback*, *sucky*, *crappy*, *nerd**, *LMAO*, *sweetie** and *scary* (the star indicates variable ending in LIWC, *e.g.*, *geek** will include *geeky, geeks*, and so on). The respective top-10 words for WNA are *dysphor*, *lachrymos*, *schadenfreud*, *scarili*, *yucki*, *horrif*, *scari*, *peski*, *gai*, *flummox* (notice the WNA terms are all stemmed). This observation seems to suggest that LIWC might be more biased towards contemporary words, which explains the robust increase present from the 1980s.

Regarding specific emotions, within the general decrease, 'Anger' and 'Digust' are the WNA categories with the highest and the lowest z-scores in 2000 respectively. This partly differs from the results presented in [7], where we identified 'Fear' and 'Disgust' as the highest and lowest trending emotions.

Finally, we consider the difference between American English and British English books (Figure 4). Again, the results obtained using WNA and LIWC are strongly correlated (Pearson's $r = 0.87$, $p < 0.0005$), and confirm that, since about 1960, American books show an increase in the usage of emotion-related terms when compared to books written in British English.

B. *Evaluation of statistical robustness*

As described in Section III-C, we apply bootstrap analysis to compute CIs for our main results in order to assess the robustness of each pattern. Figures 5 and 6 show CIs for the first main result, where we try to quantify the degree of 'Happiness' in books (Joy minus Sadness for WNA and Positive Emotions minus Sadness for LIWC). By observing the 95% CIs for WNA in Figure 5, we conclude that the mean

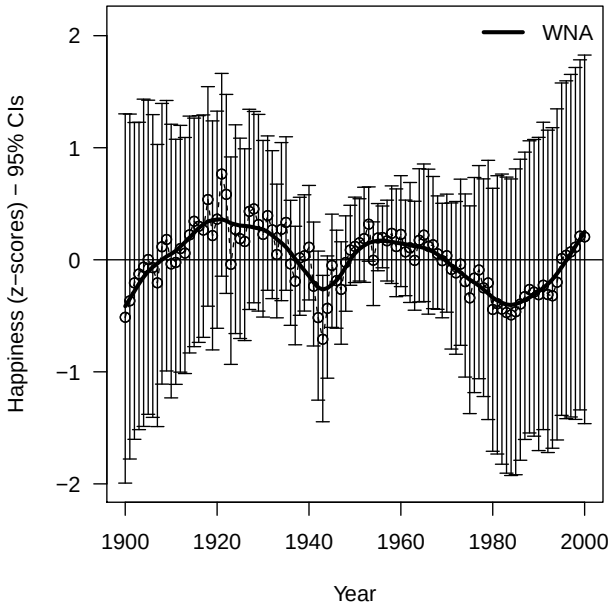

Fig. 5: 95% CIs for WNA's 'Happiness' index after applying bootstrap sampling. Circles represent the bootstrap mean and the line is a smoothed version of the bootstrap mean time series.

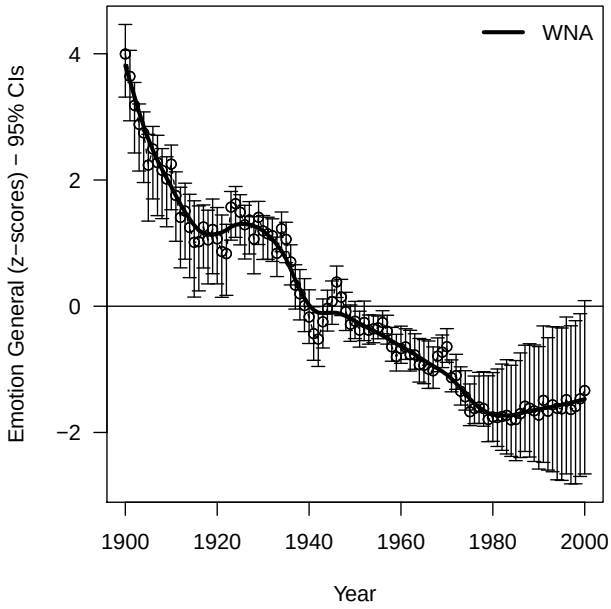

Fig. 7: 95% confidence intervals for the time series for all emotion types in WNA. Circles represent the bootstrap mean and the line is a smoothed version of the bootstrap mean time series.

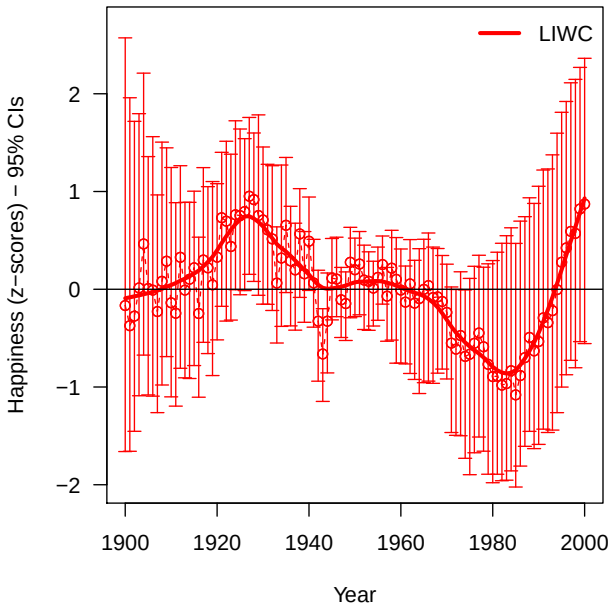

Fig. 6: 95% CIs for LIWC's 'Happiness' index after applying bootstrap sampling. Circles represent the bootstrap mean and the line is a smoothed version of the bootstrap mean time series.

signal is more consistent during a period in the middle of the century (roughly from 1920s to 1970s), but it is not as stable otherwise. To acquire a more clear picture, we repeated the same analysis for LIWC (Figure 6). In general, the 95% CIs for LIWC's 'Happiness' indicate an overall stronger robustness than WNA. However, they also show a few instability signs aligning with the WNA result (during the beginning and end of the century); still, those signs are of a definitely smaller magnitude. The means of the bootstrap samples for LIWC and WNA yearly 'Happiness' scores correlate (Pearson's $r = 0.78$, $p < 0.0005$) proving that both emotion detection tools extract a similar pattern.

Figure 7 shows the 95% CIs corresponding to the second result (b) from Figure 3, which was the decrease in emotion word use over the century [7]. Figure 8 then shows 95% CIs for the third result (c) from Figure 4, which was the divergence between American and British English books [7]. In both Figures 7 and 8, we present results only for WNA, as these were strongly correlated with the results from LIWC. Figures 7 and 8 confirm that the patterns (b) and (c) are stable across the entire century. Similarly to Figures 5 and 6, the CIs are slightly increasing during the last two decades.

C. *Part-Of-Speech analysis*

We also computed the 'Happiness' indices using the same method but limiting the data only to terms tagged as adjectives or adverbs (Figure 9 for WNA; Figure 10 for LIWC). While both have a significant linear correlation with the time series, where all terms – regardless of their POS tag – were used (for WNA: Pearson's $r = 0.66$, $p < 0.0005$; for LIWC: Pearson's $r = 0.78$, $p < 0.0005$), at visual inspection the trends appear to be qualitatively different (*i.e.*, when compared to Figure 1).

A major divergence, for example, concerns the last period of the trends, which is generally 'happier' in the original

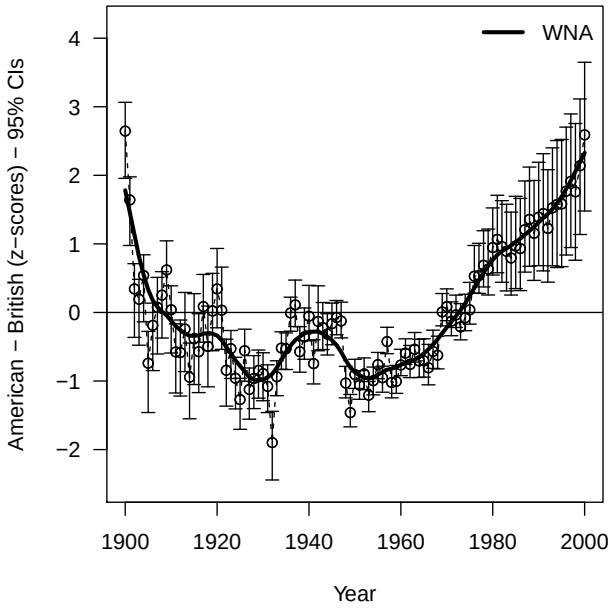

Fig. 8: 95% confidence intervals for the difference of the emotion scores in books written in American English and British English. All six emotions types in WNA are used. Circles represent the bootstrap mean and the line is a smoothed version of the bootstrap mean time series.

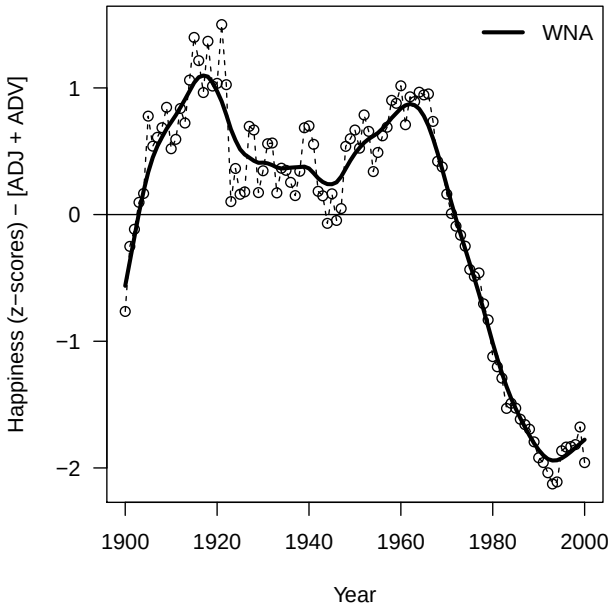

Fig. 9: z-scores of WNA 'Happiness' for years 1900 to 2000 using only adjectives or adverbs. Circles show actual times series, the line shows the smoothed time series.

analysis. Interestingly, a great part of this anomaly, in particular for WNA, can be explained by a single high-frequency term, which is the word "like". The Google Ngram corpus shows

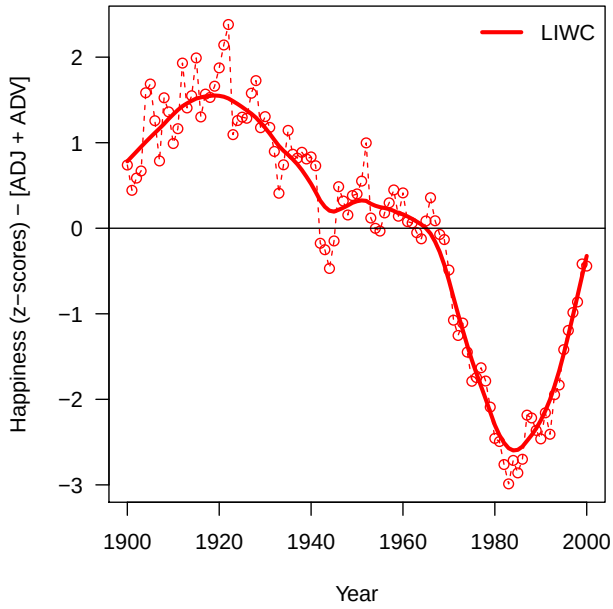

Fig. 10: z-scores of LIWC 'Happiness' for years 1900 to 2000 using only adjectives or adverbs. Circles show actual times series, the line shows the smoothed time series.

that "like" increased by 64% in frequency from 1960 to 2000 (from 0.0011 to 0.0018; both frequencies are normalized with the yearly count of "the"). This is a good example of how 'Zipf's law' [22] affects results: a single word greatly contributes to the emotion-score increase for the categories of Joy (WNA) and, to a lesser extent, Positive Emotions (LIWC). When POS is taken into account, verbs are excluded and the increase disappears, since the use of "like" as an adverb or adjective remained, according to the data in Google Ngram corpus, basically unvaried. Indeed, if we compare the original WNA 'Happiness' trend excluding the word "like" with the POS version, their correlation is sensibly higher (Pearson's $r = 0.93$, $p < 0.0005$) compared to the one when term "like" is present (Pearson's $r = 0.66$, $p < 0.0005$).

The adjective-adverb analysis reconfirms the decrease in the usage of emotion-related terms (see Figure 11; Pearson's $r = 0.99$, $p < 0.0005$ with the WNA pattern depicted in Figure 3) as well as the divergence between books written in American and British English (see Figure 12; Pearson's $r = 0.97$, $p < 0.0005$ with the WNA pattern depicted in Figure 4). All results from the POS analysis are in line with the robustness levels indicated by the CIs and presented in the previous Section, *i.e.*, WNA and LIWC 'Happiness' indices are less stable than the other emotional patterns.

V. Conclusions and Future Work

We overall find a substantial agreement between the tools we used to measure emotional word frequencies in publications over the 20$^{\text{th}}$ century. There were some notable specific differences, however.

First, while the general 'Happiness' scores of WNA and LIWC were correlated, their trends substantially differed from

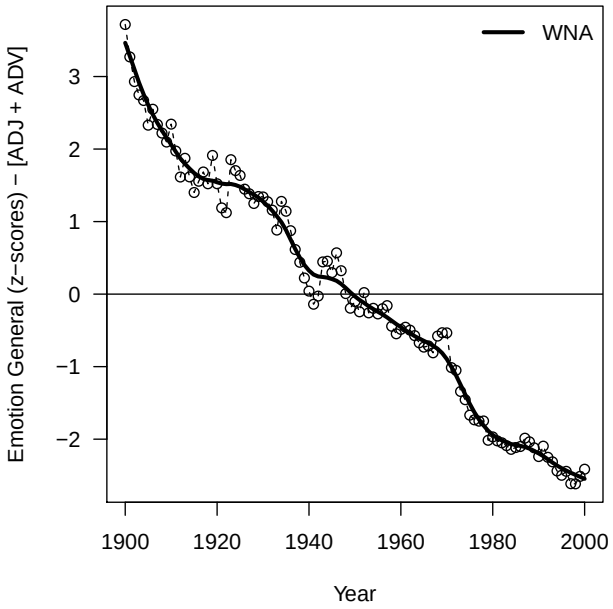

Fig. 11: Re-analysis of WNA (see Figure 3) using only terms tagged as adjectives or adverbs. Plot shows the difference between z-scores of the six emotions versus a random sample of stems from 1900 to 2000. Circles show actual times series, the line shows the smoothed time series.

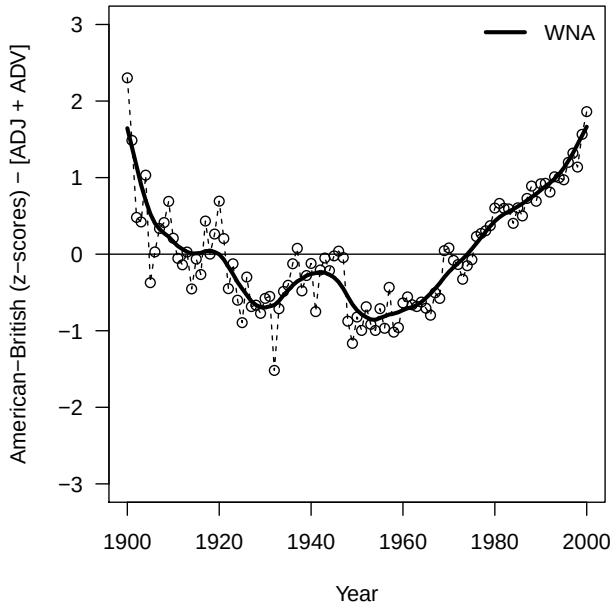

Fig. 12: Difference, using only terms tagged as adjectives or adverbs, between z-scores for emotion terms in American English and British English for years from 1900 to 2000. Aggregate score of the six emotions with WNA. Circles show actual times series, the line shows the smoothed time series.

the one retrieved using HED. The reason for this may rely on the fact that HED is conceptually different from both WNA and LIWC, as it contains high-frequency terms evaluated for their degree of 'Happiness', and not terms associated to particular emotions.

Secondly, we found a quite different trend in aggregated frequencies of all emotional words, from about 1980 to 2000, between WNA and LIWC. This may be related to differences between the contents of each emotional taxonomy; in fact, the sets containing the top-10 terms with a relative frequency increase during this period for LIWC and WNA were completely disjoint.

Finally, a clear divergence was evident when we considered only adjectives and adverbs, rather than all emotion words. Imposing those constraints on the input data increased the variance between WNA and LIWC results and also yielded a qualitatively different pattern in the case of literary 'Happiness'. We found again, however, that this anomaly might be largely explained by isolated specific words. For example, simply removing one word ("like") from the comparison caused the non-POS and POS WNA results to fall back to similarity with each other.

In general, the patterns of literary 'Happiness' were proven less robust considering the statistical proof (bootstrap CIs) and the additional POS analysis; hence, further investigation may be essential. On the contrary, the decline of emotion-word usage and the divergence between American and British English books were consistent across our analysis.

In a broader perspective, intuitively expected results – for example, the fact that books tended to have a lower 'Happiness' score during Second World War – may provide some confidence that more surprising results, such as the the decline of emotion-word usage, and the divergence between American and English books, represent real patterns.

Future challenges could include the incorporation of n-grams with $n > 1$ as they have a clearer semantic interpretation (*e.g.*, "like" in the expression "I don't like X" is considered as a positive emotion term in our current analysis) as well as the investigation of languages other than English. Studying a variety of different human-generated inputs (blogs, Social Media, news articles, but also movie scripts and TV/radio transcripts) could also reveal interesting patterns. For example, a quick comparison between the non z-scored time series of HED (right y-axis in Figure 2) with scores obtained from recent (2008-2011) Twitter data [17], reveals that books appear to be on average less 'happy' (with values ranging, during the century, approx. between 5.9 and 5.95, whereas Twitter content is often scored > 6).

In conclusion, the study of cultural evolution [29]–[31] might greatly benefit from the current availability and abundance of quantitative data. While, as for any new field of research, much work is needed to assess its full potential and applicability, our overall results are encouraging in regard to the application of such methodologies for studying long-term cultural dynamics.

ACKNOWLEDGMENTS

Alberto Acerbi is supported by the British Academy and the Royal Society through a Newton International Fellowship. Vasileios Lampos acknowledges the support from the Trend-Miner project (EU-FP7-ICT n.287863). All authors would like to thank J. Pennebaker for providing the lists of emotion-related words from LIWC-2007.

REFERENCES

[1] S. Arbesman, "Stop hyping big data and start paying attention to 'Long data'," http://www.wired.com/opinion/2013/01/forget-big-data-think-long-data/, 2013.

[2] J. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, pp. 176–182, 2011.

[3] P. S. Dodds and C. M. Danforth, "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents," *Journal of Happiness Studies*, vol. 11, no. 4, p. 441–456, 2009.

[4] C. N. DeWall, R. S. Pond Jr, W. K. Campbell, and J. M. Twenge, "Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular us song lyrics." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 5, no. 3, p. 200, 2011.

[5] J. M. Twenge, W. K. Campbell, and B. Gentile, "Increases in individualistic words and phrases in american books, 1960–2008," *PLoS ONE*, vol. 7, no. 7, p. e40181, 2012.

[6] S. Mohammad, "From once upon a time to happily ever after: tracking emotions in novels and fairy tales," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, p. 105–114.

[7] A. Acerbi, V. Lampos, P. Garnett, and R. A. Bentley, "The expression of emotions in 20th century books," *PLoS ONE*, vol. 8, no. 3, p. e59030, 2013.

[8] Y. Lin, J. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, p. 169–174.

[9] G. A. Miller, "WordNet: a lexical database for english," *Communications of the ACM*, vol. 38, p. 39–41, 1995.

[10] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," in *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, p. 1083–1086.

[11] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, ser. SAC '08. New York, NY, USA: ACM, 2008, p. 1556–1560.

[12] V. Lampos, "Detecting events and patterns in Large-Scale user generated textual streams with statistical learning methods," arXiv e-print, 2012. [Online]. Available: http://arxiv.org/abs/1208.2873

[13] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Nowcasting the mood of the nation," *Significance*, vol. 9, no. 4, pp. 26–28, 2012.

[14] V. Lampos, T. Lansdall-Welfare, R. Araya, and N. Cristianini, "Analysing mood patterns in the united kingdom through twitter content," arXiv e-print, 2013. [Online]. Available: http://arxiv.org/abs/1304.5507

[15] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count (LIWC2007)," Tech. Rep., 2007.

[16] J. W. Pennebaker, *The secret life of pronouns: what our words say about us*. New York: Bloomsbury Press, 2013.

[17] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PLoS ONE*, vol. 6, no. 12, p. e26752, 2011.

[18] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, 2010, pp. 178–185.

[19] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[20] M. Perc, "Evolution of the most common english words and phrases over the centuries," *Journal of The Royal Society Interface*, 2012.

[21] M. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[22] G. K. Zipf, *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley, 1949.

[23] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, p. 1–135, 2008.

[24] M. F. Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130–137, 1980.

[25] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore, "Quantitative patterns of stylistic influence in the evolution of literature," *Proceedings of the National Academy of Sciences*, vol. 109, no. 20, pp. 7682–7686, 2012.

[26] R. A. Bentley, P. Garnett, M. J. O'Brien, and W. A. Brock, "Word diffusion and climate science," *PLoS ONE*, vol. 7, no. 11, p. e47966, 2012.

[27] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. London UK: Chapman & Hall/CRC, 1993.

[28] J. H. Friedman, "Smart user's guide," Laboratory for Computational Statistics, Stanford University, Palo Alto, Tech. Rep., 1984.

[29] R. A. Bentley, M. W. Hahn, and S. J. Shennan, "Random drift and culture change," *Proceedings of the Royal Society B: Biological Sciences*, vol. 271, no. 1547, pp. 1443–1450, 2004.

[30] A. Acerbi, M. Enquist, and S. Ghirlanda, "Cultural evolution and individual development of openness and conservatism," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 931–18 935, 2009.

[31] A. Acerbi, S. Ghirlanda, and M. Enquist, "The logic of fashion cycles," *PLoS ONE*, vol. 7, no. 3, p. e32541, 2012.