Multi-lingual Valence Analysis Across 20th Century Literature and the Twittersphere

A Thesis Presented

by

Eric M. Clark

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fullfillment of the Requirements
for the Degree of Master of Science
Specializing in Mathematics

May, 2014

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Master of Science, specializing in Mathematics.

Thesis Examination Committee:

Peter Dodds, Ph.D. — Advisor

Chris Danforth, Ph.D.

Jacques Bailly, Ph.D. — Chairperson

Cynthia J. Forehand, Ph.D. — Dean, Graduate College

Date: March 24, 2014

Abstract

Understanding and statistically processing underlying trends in natural human language has been an ongoing goal in Computational Social Science. This work explores trends in several languages, using expressions found on the internet, in 20th century literature, and social media. We use a Hedonometer to measure happiness in several corpora, using human ratings of emotionally charged words. Previous work has established and tested the instrument on English corpora, discovering a bias towards positive word usage in billions of tweets, millions of books, music lyrics, and media articles. Until now, it has remained an open question as to whether this trend is prevalent with respect to other languages. This work extends these previous analyses through a multilingual extension of the hedonometer to uncover interesting stories and underlying trends from literature and across social media.

I dedicate this thesis to my family for their endless love support and encouragement.

Acknowledgements

I would like to sincerely thank my advisors Professor Peter Dodds and Professor Christopher Danforth for their support and guidance throughout this study and during my studies as a graduate student at the University of Vermont. I would also like to thank the members of the Computational Storylab for their invaluable insight and advice. I also wish to acknowledge the Vermont Advanced Computing Core, which is supported by NASA at the University of Vermont which provided High Performance Computing resources along with the MITRE Corporation that contributed to the research results reported within this thesis.

To all my friends, thank you for your support and understanding during my times of absense due to academic commitments. Finally I'd like to leave the remaining space in memory of Harry Bakalian (1921-2011) a loving father, grandfather, and war veteran whose wisdom helped shape me into the scholar I am today.

Contents

List of Figures

Chapter 1

Introduction and Literature Review

Various tools have been established for the analysis of sentiment in written text. This work focuses primarily on measuring the contribution of emotional words used in a corpus, a methodology known as hedonometrics (Dodds and Danforth 2009). Hedonometrics was developed and thoroughly tested on several English corpora in (Dodds and Danforth 2009), (Dodds et al. 2011), (Kloumann et al. 2012). This work extends these analyses to several other languages with the aim of uncovering trends across literature and a popular social media outlet known as twitter. Using frequency distributions of words appearing across 20th century literature, significant historical events correspond to time periods with an abundance of strongly emotional literature. Applying this metric to the twittersphere can identify small-scale events occurring internationally to within an hourly time scale.

1.1 Introduction

The natural emotions people tie to words can be used as a powerful tool to uncover underlying trends and hidden stories when analyzing natural expressions in literature and across social media. (Dodds et al. 2011) have created a measure to numerically classify the emotional content of text known as the hedonometer. Running hedonometric analysis on social media can elucidate events in real time. When restricting to literature, important historical events correlate to timeperiods that use extremely emotional language. (Dodds et al. 2011) centered its analysis across English lexica. This work extends this analysis to include several foreign languages to uncover stories across social media and 20th century literature.

1.2 Validating the Pollyanna Hypothesis Across Language

The Pollyanna Hypothesis, (Boucher and Osgood 1969), claims the existence of a subconscious human tendency to recall optimistic rather than pessimistic events in daily life. (Kloumann et al. 2012) has shown that this hypothesis is linguistically valid over various electronic media, English literature, and song lyrics. However, it remains an open question whether this result holds multilingually. In an effort to resolve the issue, we reproduce the methods of (Kloumann et al. 2012) with multilingual corpora, observing a positivity bias across written human expression.

1.2.1 Word List Distributions

To extend (Kloumann et al. 2012), data driven word lists are compiled for the most frequently used words appearing in French, Spanish, German, Korean, Portuguese, and Russian text. Multiple corpora were sampled to ensure each list had a robust yet unbiased sampling of written human expression. Each wordlist distribution was composed from three primary corpora: Google Books, The Google Web Crawl, and Twitter.

Google has been scanning and digitizing millions of books to create a free online database that's readily available to the public. The google books n-gram data-set is a free online database of the frequency distributions of all n-grams (phrases of n-words) occurring from 1505 to 2008, (Michel et al. 2010, Google-Labs). Each of these distributions has been categorized by language and includes the year and frequency of each n-gram that has appeared at least forty times. Web 1T, also known as the Google Web Crawl (Brants and Franz 2006), implemented tokenization software that compiled the most frequently occurring words on the internet. The Google Web crawl contains over a trillion tokens that have been collected from public web pages. (Twitter) is a popular social networking tool that allows users to interact via short 140 character messages called tweets. The twitter frequency distribution was compiled from the words (characters between a space) appearing from a full year of tweets. Each of these frequency distributions were combined to create language specific wordlists of the most frequently occurring 10,000 words per language.

Native speakers from each language were recruited from (Appen-Butler-Hill), a company that outsources translation of call-center audio transcripts to an international network of tens of thousands of bilingual individuals. People were asked to rate the emotional responses elicited by words appearing on each distribution. Surveys implemented a pictorial scale comparable to the self affective mannequin (SAM) developed in (Bradley and

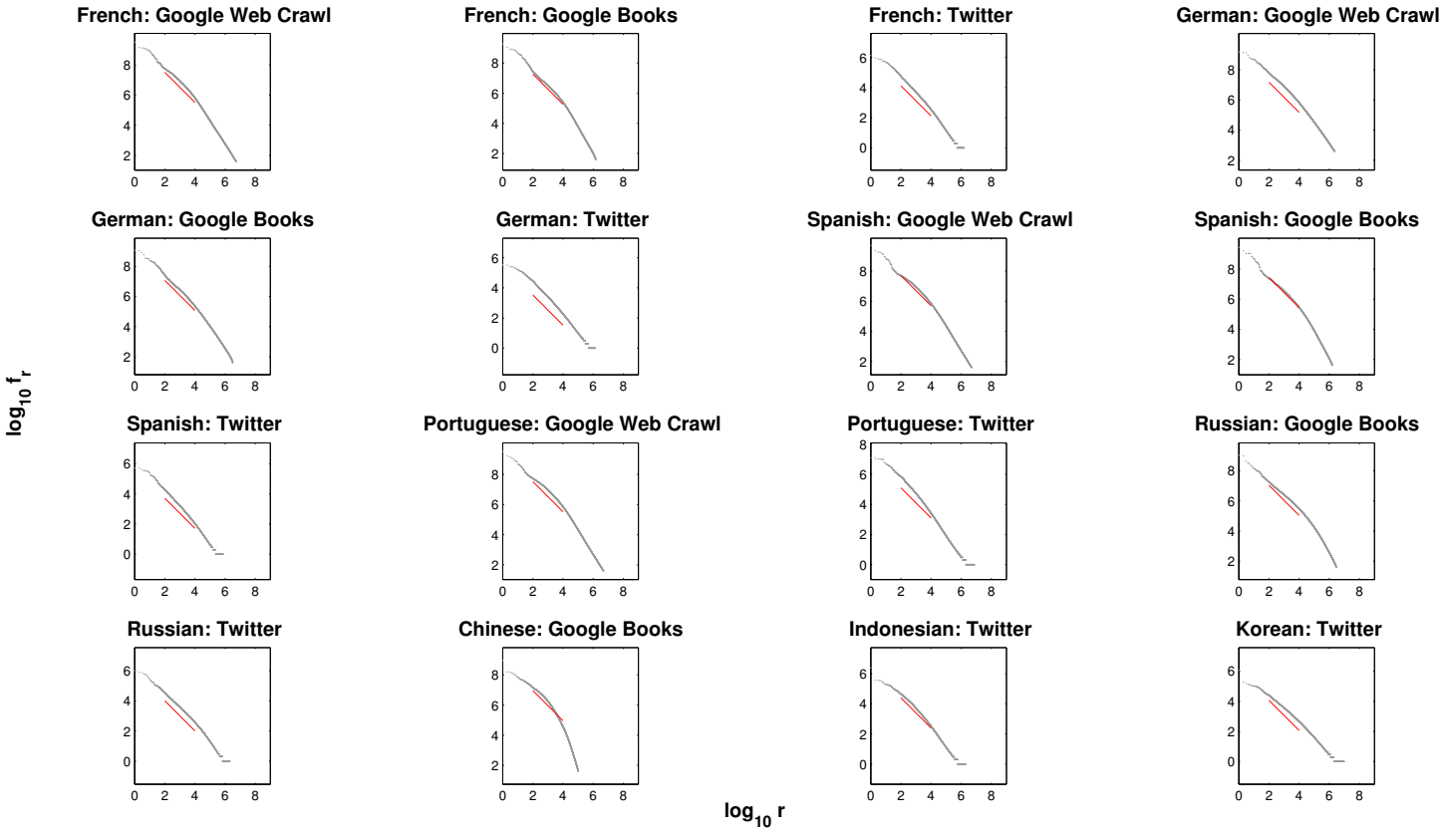
Lang 1994). This adaptation of SAM incorporates faces ranging from frown to a smile. For each word, participants were directed to choose the face that best represented the appropriate emotional response. The responses were then converted to a 9 point scale. On the numeric scale, 1 corresponded to the face with the largest frown and 9 to the face with the largest smile. The average happiness score, h_{avg}, for each word is then calculated via the arithmetic mean of 50 user reported ratings per word. The finished product is a data driven list of the most frequently abundant 10,000 words appearing in each language with a corresponding average happiness value.

The "Language assessment by Mechanical Turk", or LabMT for short, compiled by (Dodds et al. 2011) is implemented to conduct the analysis on English corpora. This distribution was rated using an identical survey structure described above. Participants were employed via a popular online survey tool, (Amazon-Mechanical-Turk), to rate each English word. English words with low ratings ($h_{avg} < 4$) are defined as negative words, examples of which include 'kill', and 'terrorist'. Positively charged words have a rating, $h_{avg} > 6$ and include 'laughter', 'happy', and, 'love'. Neutral words are defined as words in the range $4 < h_{avg} < 6$ examples include 'and', 'the', 'of', and 'it'.

For large literary corpora, Zipf's law states that the frequency of English words, $f(r)$, are inversely proportional to their rank, r, (Zipf 1949). Mathematically, this relationship is represented as $f(r) = r^{-\alpha}$ where $\alpha \approx -1$ has been verified for an abundance of English corpora. This relationship is approximately linear when plotted on a logarithmic scale. (Petersen et al. 2012) has shown a similar relationship holds for several other languages. Figure 1.1 plots the rank of each wordlist against its frequency on a logarithmic scale alongside a line with slope -1 as a reference to demonstrate the nearly Zipfian quality of each corpus. Due to Zipf's law, a significant density of all text is governed by the highest

ranked words. Figure 2.1 displays the highest 5 ranking words for several of these wordlist frequency distributions. In each language these are identified as articles which are rated as neutral words. Figure 1.1 demonstrates the overpowering frequency contribution of these neutral words. For significantly large textual analysis, the overabundance of these articles will dampen the signal of emotionally charged words. The emotional signal however is not extinguished as shown in (Dodds et al. 2011), hence it's advantageous to exclude neutral stop words ($4 < h_{avg} < 6$) in order to bolster the emotional signal of literature and twitter data.

Figure 1.1: Wordlist Distributions



1.2.2 Translating Happiness Multilingually

Properly translating large corpora is a challenging linguistic task. Nevertheless, we make an attempt here to compare concepts of an emotional score to translated versions. The Affective Norms for English Words (or ANEW) presented in (Lang 1999) is another English word distribution that incorporates the SAM survey technique to rate approximately 1000 words over several emotional spectrums including happiness. Both (Redondo et al. 2007) and (Soares et al. 2012) have shown high correlations between the user reported valence of Spanish, English, and Portuguese translations for the ANEW dataset. Google's free online translator, (Google-Translate), allows for a lexical comparison of happiness scores between each wordlist. It should not be a surprise that the word 'happy' preserves its high positive rating under translation. Not all words have such an explicit lexical preservation post translation. For instance the word 'matt' in English is rated as neutral, but the online translation to Arabic is an extremely negative word that is commonly indicative of 'death'. These outlier translations can skew the interpretation of inter-language valence analysis. Hence employing native speakers to complete happiness surveys is crucial for a rigorous reproduction of (Kloumann et al. 2012).

1.3 Happiness Timeseries and Wordshift Graphs

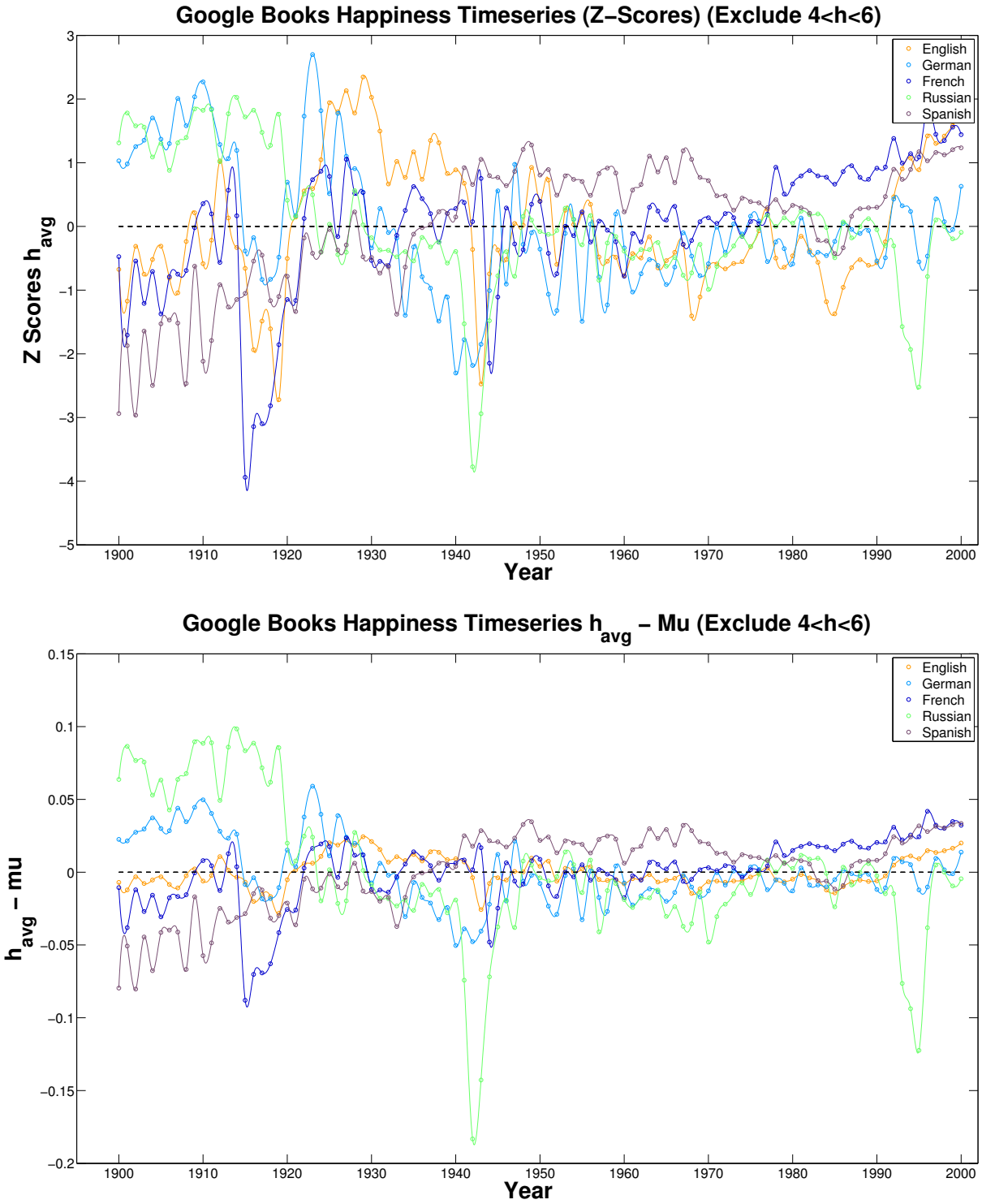
Using the happiness scores of each wordlist distribution, the average emotional rating of a corpus is calculated by tallying the appearance of words found in the intersection of the wordlist and a given corpus. A weighted arithmetic mean of each word's frequency, f_{word}, and corresponding happiness score, h_{word} for each of the N words in a text yields

the average happiness score for the corpus, $\bar{h}_{text}$:

$$\bar{h}_{text} = \frac{\sum_{w=1}^{N} f_w \cdot h_w}{\sum_{w=1}^{N} f_w}$$

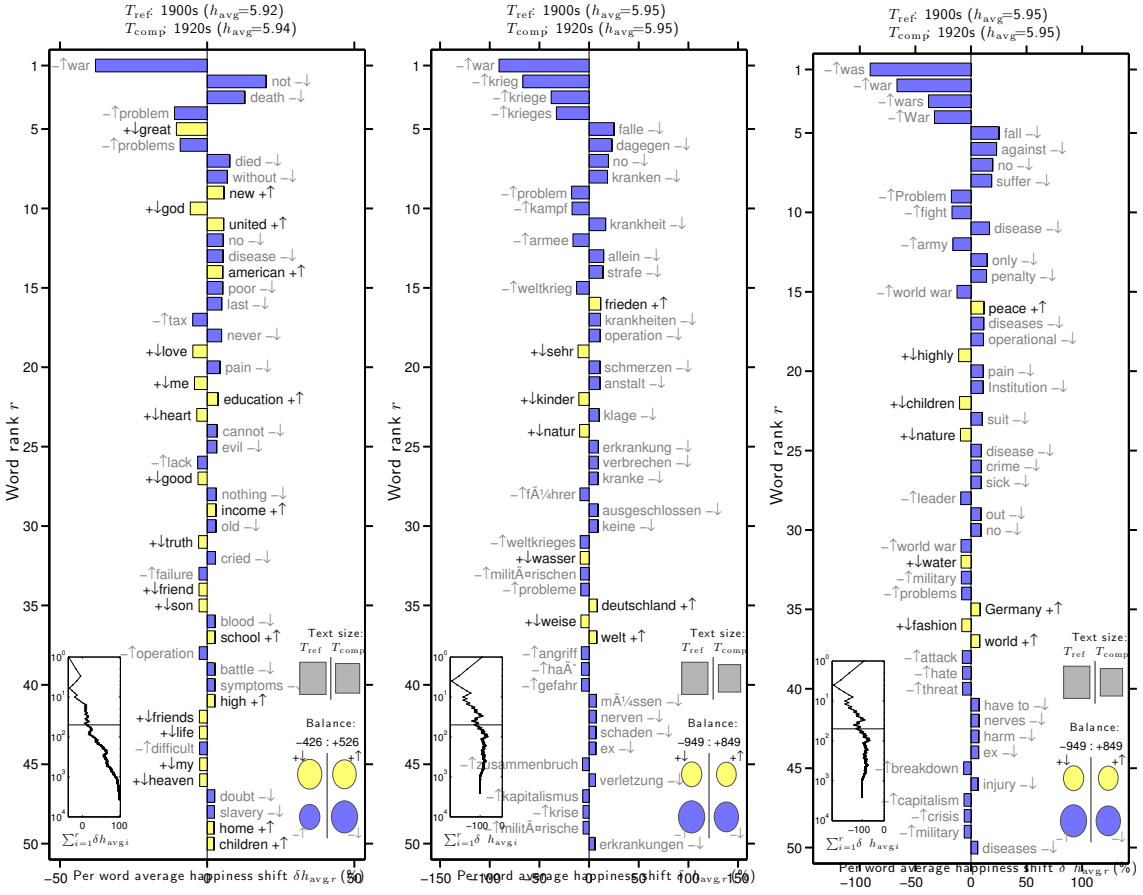
Happiness time-series are created for corpora that provide time stamps with a corresponding word frequency, to illustrate periods that use an abundance of emotional text. (Dodds et al. 2011) have shown that significant events are identified as outliers of the happiness time-series of twitter data. (Acerbi et al. 2013) conducted an English valence analysis over Google Books dataset using the LabMT happiness distribution. This analysis can now be extended by incorporating each multi-lingual wordlist distribution. Figure 1.2 shows a happiness time series of 20th century literature for several languages. In the spirit of (Acerbi et al. 2013), each data point on the upper plot corresponds to the usual Z-score (see methods section) of the average happiness value, excluding neutral stop-words ($4 < h_{avg} < 6$). For comparison, the lower plot displays the average happiness per year with the global mean subtracted from each curve. The same trends are present in both plots, however the Z-scores help indicate time periods that are significantly more emotional in terms of the global standard deviation. These data points were connected using cubic spline interpolation techniques to provide a smooth representation of happiness transitions between each year. The outliers on the time-series plots for 20th century literature are indicative of significant events in history. For example, both world wars can be identified on the plot as outliers. After identifying these interesting time-frames, its important to understand the words that are contributing to the change in average happiness.

Figure 1.2: Google Books Happiness Time-series



To provide a thorough analysis of the emotional trends of a corpus, it's important to identify the key words that have the most significant impact in creating positive or negative shifts in average happiness. (Dodds and Danforth 2009) have created word-shift graphs that depict the English words responsible for a shift in average happiness when compared to the word frequencies of a specified reference period. This tool has been modified for use with the multi-lingual happiness distributions. For reference, the English translations of words that appear on these new word-shifts are portrayed side by side. Translations were obtained through Google's free online translator and thus are not perfect. However, these translated word shifts yield quick insight to the types of words that are responsible for the multi-lingual emotional shifts.

Figure 1.3: English vs German Google Books Word-shift Graph



Word-shift graphs illustrate two separate word frequency distributions. A reference period (T_{ref}), creates a basis of the emotional words being used to compare with another period, (T_{comp}). The top 50 words responsible for a happiness shift between the two periods are displayed, along with their contribution to effecting the average happiness of the era. A few example word shift graphs are presented in Figure 1.3. The graph farthest to the left represents the different types of words appearing over two separate decades from the English Google Books corpus. The reference period word distribution spans from 1900 to 1910 while the comparison word distribution spans from 1920 to 1930. The arrows ($\uparrow, \downarrow$) next to a word indicate an increase or decrease, respectively, of the word's frequency during the comparison period with respect to the reference period. The addition and subtraction signs indicate if the word contributes positively or negatively, respectively, to the average happiness score. For example, the word that has the greatest negative contribution to happiness is the word 'war' at the top of the list. The up arrow indicates "war" was used more frequently during the 1920-30s and the - sign indicates this has a negative contribution on the average happiness score for this time period.

The subgraphs within each wordshift display information regarding the reference and comparison corpora. The "Text Size" subgraph displays the relative frequency size of each corpora. Since each of these corpora are taken from a full decade of literature data, their frequencies are also comparable, so the squares that represent the size of each corpora have a similar size. The "Balance" subgraph represents the abundance of words that are contributing to positive or negative shifts in valence as circles. For example, the yellow circle on the upper left is an indication of the positive words that are less frequent in the comparison text. The left-most subgraph displays the happiness weight of all of the words in the corpus. The line represents the cutoff between the words being displayed on the

wordshift and the words that have been omitted. This yields information regarding the entire corpus that is not represented by the top 50 ranking words appearing on the wordshift.

The word-shift plots in the center and far right of the above figure are from the German Google Books corpus centered around the same time period. While the average happiness for each decade is quite close (both show 5.95) ther is a general agreement with the themes observed in English. The central graph are the raw German words that have a significant contribution in displacing the average happiness of the comparison period with respect to the reference period. The graph on the far right are the words from the central graph translated to English using Google's free online translation service. This allows a non-German speaker to understand the types of words being used in the German corpus. Word-shift plots yield insight into the literary emotions that have been encoded in the Google Books corpus as a consequence of significant historical events.

1.4 Google Books: Hedono-History Tour

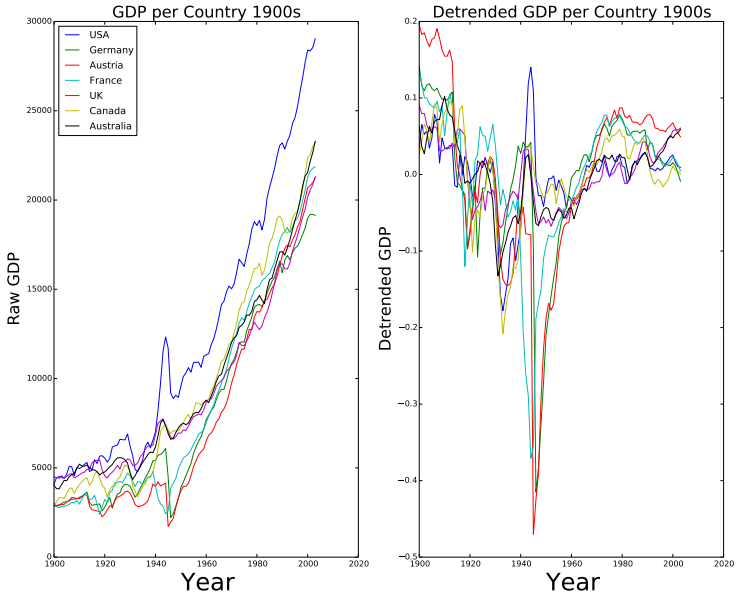
The Google books n-gram dataset includes each word's frequency of occurrence per year of appearance. This allows for a multi-lingual time-series valence analysis across the google books corpus. Large scale historical events can be identified as local optima of the average happiness time-series. However, to infer that a specific time-period is connected to a trend in the time-series analysis, the specific words that influence the negative or positive shift must be identified. Word-shift graphs, centered about extreme points in the time-series, project the literary mood of the time-period and can reflect times of war, struggle, or prosperity. For example, the time leading up to and directly after both world wars has a significantly lower valence than other more peaceful time periods. Word-shifts centered around these time periods help identify these global events and their impacts on literature.

1.4.1 20th Century Per Capita GDP and Hedonometric Connection

Understanding the relationship between the literature of a nation and its economic standing has practical potential. (Bentley et al. 2014) has demonstrated that the emotions present in literature correlate to the economic misery of the recent past for both the US and Germany. "Economic misery" of a country is defined as the sum of the inflation rate and unemployment rate, which yield insight to the quality of life of a nation's constituents. Their analysis reports the existence of an optimal lag between the emotional state of a time period and the economic misery of the nation.

The Gross Domestic Product (or GDP) is a measure of the total value of all goods and services produced within a given country. Per capita GDP is a ratio of the GDP and population of a country. This measurement can also be used as an indicator for the quality of life within a nation and is used to compare the economic performances of different countries (Easterlin and Angelescu 2009). Since per capita GDP reflects a nation's quality of life, it's reasonable to predict a similar connection to the emotional trends in literature. However, as noted in (Easterlin 1974), the GDP of two nations can not accurately distinguish the reported happiness difference between the nation's citizens. Hence it is unfounded to assume that nations with a substantially higher GDP per capita will necessarily have a happier literary mood. GDP per capita data for several countries over the 20th century was downloaded from a free online database, (The-World-Bank). Proceeding in the same manner as (Bentley et al. 2014), an optimal lagging relationship is identified between GDP per capita and its effect on the emotional response of literature.

Figure 1.4: GDP per capita 20th Century



GDP, along with the population of each country, has been steadily increasing throughout the 20th century. This forces the GDP per capita of each country in this analysis to steadily exponentially increase. Correlating the raw GDP data with raw average happiness of literature over time could lead to a misleading high correlation that is not a fair representation of the relationship between these two variables. To prevent a false-result, word happiness ratings between 4 and 6, designated as ”stop words”, are excluded from the analysis. Since the GDP per capita exhibits exponential growth, the dataset is transformed by the base 10 logarithm. The transformed dataset is de-trended by removing the best linear fit regression line. Shifts in the de-trended GDP per capita data set are more indicative of the relative economic change per year of each country. By cleaning the data in this fashion, a high correlation between these two variables is a much stronger result.

1.5 Twitter Time Series Analysis

Online blogging has become a popular outlet for a wide range of demographics to express their opinions on current events (Lenhart and Fox 2006). The Complex Systems Center at UVM has been collecting approximately 10% of all twitter data for the past five years. Over this time period the data feed went from collecting approximately half a million tweets per day to a staggering 50 million tweets per day; a social media explosion that has been compiled into an invaluable data set of staggering proportions. Since the Google books corpus only provides time stamps on the annual level, only large scale historical signals are observable. Twitter publishes time stamps accurate to the order of seconds, which can pick up the signal of significant events at a previously unresolveable temporal scale. As found in (Dodds et al. 2011), dips and peaks of the twitter happiness time series generally correspond to interesting real world phenomena.

The Olympics serve as a testbed to compare a universally emotional time-period world-wide. The Olympics are celebrated by a wide variety of nations whose combined twitter-active constituents yield significant coverage for each of the new multi-lingual valence wordlists. Worldwide popularity for this extremely emotional time-period allows for the opportunity to apply each multi-lingual happiness distribution to recover international emotional responses of the outcome of the olympic games. The specific times at which olympic medals were awarded were compiled from official media reports, (New-York-Times).

Chapter 2

Methods

2.1 Tokenizer: Creating Wordlist Distributions

A 'tokenizer' is a tool that breaks a text down into a predefined list of acceptable symbols, known as tokens, by removing special characters and then creating a frequency distribution of the number of times each token appears in the text. Here, the tokens are representative of common words and are created by parsing the text by space and then removing all punctuation occurring before and after the token. This creates a list of the most commonly used words in each corpora. Perl, a popular programming language, has a built-in regular expression pattern matching software that was implemented to match tokens in each corpora and create frequency word distributions for Google Books, the Google Web Crawl, and tweets. Words were selected from each list and compiled into a final list consisting of the 10,000 most frequently used words.

Figure 2.1: Most Frequently Used Words Per Language

Rank	Spanish	Portuguese	Indonesian	French	German	Russian	Chinese
1	que 5.10 (that)	que 5.00 (that)	d 4.86 (d)	de 5.06 (of)	ich 6.20 (I)	в 5.00 (in)	的 5.10 (of)
2	de 5.42 (of)	de 5.30 (of)	yg 4.88 (reply)	la 5.30 (the)	und 5.26 (and)	и 5.08 (and)	是 5.42 (be)
3	a 5.26 (to)	e 5.08 (and)	di 5.12 (in)	je 5.40 (I)	die 4.82 (the)	не 4.42 (not)	在 5.02 (in)
4	y 5.34 (and)	o 4.84 (the)	aku 5.76 (I)	le 5.08 (the)	in 5.50 (in)	на 5.12 (on)	了 5.08 (the)
5	la 5.30 (the)	a 5.24 (the)	ya 6.04 (yes)	a 5.00 (has)	der 5.18 (the)	я 6.50 (I)	和 5.46 (and)

Figure 2.2: Tokenizer Example

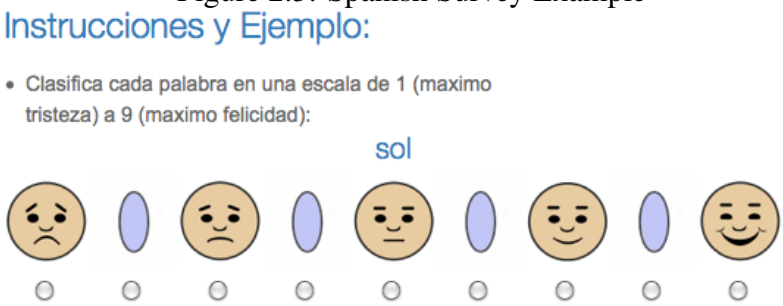
Non Tokenized List	Tokenized List
love! 2000 .love 2000 @love 1000 love87 1000 LoVE 500 love 10000	love 14500 love87 1000 (@love would be thrown away)

2.2 Surveys

Appen Butler Hill recruited native speakers to rate the words on each wordlist. For each language, native speakers are employed to identify the face in Figure 2.3 that best matches the emotional response elicited by each word. Each wordlist was compiled by taking the arithmetic mean of 50 happiness scores per word and recording the standard deviation, as

well as the rank of the word from each distribution it was compiled from. A sample from the survey is displayed below.

Figure 2.3: Spanish Survey Example



Here the Spanish word Sol is displayed. Survey participants are asked to choose the face that is the best indication of how this word effects them emotionally. This scale is then translated into a numerical 9 point happiness scale. In this example, 'sol' scored 7.66, while in English the corresponding word 'sunshine' was rated at 7.94.

2.3 Inter language Translations

Google provides a free online translation service that we used to translate each wordlist into each other language. Not all words have a bijective correspondence under the google translator. These unstable words will translate to a specific word in another language but do not translate back to the original word. Unstable translations were excluded from the comparison. The intersection between each wordlist was then compiled, along with the happiness scores corresponding to each language. Correlations between scores in each language were calculated using standard linear regression.

2.4 Google Books and Twitter Happiness Time-series

Using Perl's built in case-insensitive regular expression toolbox, the frequency of words appearing on both the corpora and each wordlist were tallied along with their time of appearance. The google books corpus reports the year of appearance while twitter provides a full timestamp for each word appearing in the dataset. The average happiness per time is recorded in a file for analysis. To incorporate different lenses, the average happiness is computed with all the data, and then excludes ranges of different stop words. For this analysis words with happiness scores between 4 and 6 were excluded from the distribution.

The Google Books happiness time-series display the standard Z-score of each year's average happiness as in (Acerbi et al. 2013). To compute the Z-score, $Z_{h_{avg}(t)}$ for any given year, t, the language specific global happiness average of the 20th century, $\mu_{h_{avg}(1900-2000)}$, is subtracted from the average happiness score for that year, $h_{avg}(t)$ and then normalized by the global happiness standard deviation, $\sigma_{h_{avg}(1900-2000)}$:

$$Z_{h_{avg}(t)} = \frac{h_{avg}(t) - \mu_{h_{avg}(1900-2000)}}{\sigma_{h_{avg}(1900-2000)}}$$

2.5 GDP per capita Analysis

GDP per capita for several countries were downloaded from a free online database, (The-World-Bank). Since both GDP and population of each country are gradually increasing over the 20th century, the dataset is de-trended by removing the best linear fit from the log transformed GDP data. To test the hypothesis that literature and GDP have a lagging correlation, the average happiness score per year of 20th century literature with a variable

time lag is correlated with the GDP per capita using Spearman and Pearson correlation coefficients. Proceeding as in (Bentley et al. 2014), both a variable time lag and a moving average window are separately implemented.

A variable time lag of t years is introduced by correlating GDP data over the years 1900 to $2000-t$ with literature average happiness between $1900+t$ and 2000. This ensures both vectors are of equal length. The moving average window of t years over the GDP data is calculated as follows. The first data point is obtained by averaging the GDP between year 1900 and $1900+t$. The second data point is obtained by averaging the points between 1901 and $1901+t$. This process is continued until the sum of year of the data point and t exceeds 2000. The resulting vector is then correlated with the happiness vector spanning $1900+t$ and 2000 years. The window for the moving average and discrete time lag are varied from $t=0$ to $t=40$ years.

Chapter 3

Results

3.1 Pollyanna Hypothesis Across Language

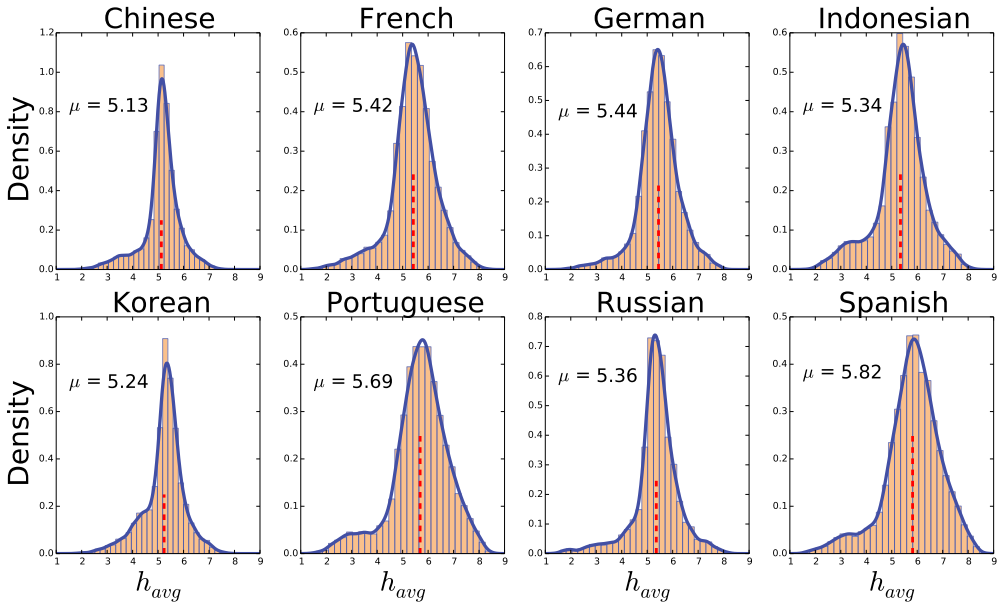
A positivity bias in the emotional scores for each wordlist, twitter data, and google books is visualized via histogram plots. Inferences on the population are visualized via kernel density estimation curves. An inherent positivity bias is present in the language prevalent across the twitter sphere, literature, and websites. Next, the correlations of happiness ratings of translated words across each language are presented via heat map scatter plots. For each wordlist distribution there is a significant correlation between the happiness scores of translationally invariant words. This is an indication that, on average, the emotional meaning of words is preserved culturally across language.

3.1.1 Wordlist Valence Distributions

Each wordlist distribution is created via frequency distributions of sources of natural written expression. Hence the happiness distributions of the wordlists provide insight to the

density of emotional words appearing across several mediums of written expression. Average happiness distributions of each wordlist are represented as histograms in Figure 3.1. Each distribution displays a positivity bias despite a fair sampling of both positive and negative words. Each distribution exhibits gaussian attributes with means between 5 and 6. These wordlist distributions serve as an unbiased tool to measure the density of emotional words on other corpora.

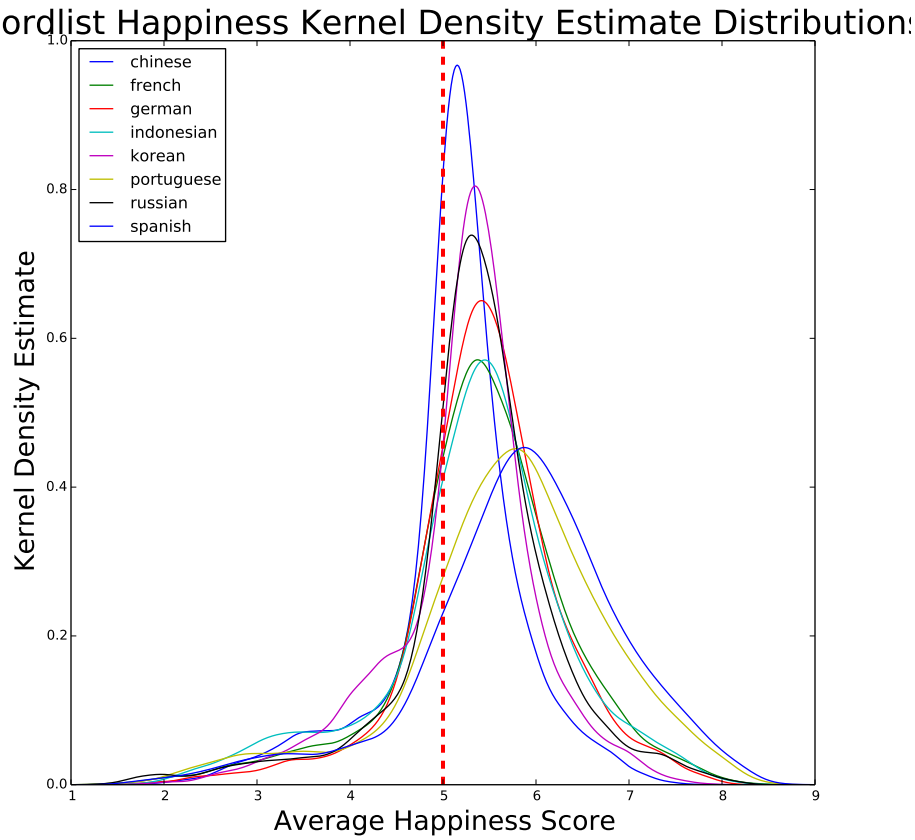
Figure 3.1: Wordlist Distributions



Using Python's built in science toolkit, kernel density estimation curves show that each wordlist is normally distributed with means that are slightly above neutral. Kernel density estimators provide insight into the global trends of each wordlist distribution. This provides a subtle indication that a higher density of positively charged words exist across literature,

websites, and Twitter. To investigate this claim, frequency weighted valence histograms of each corpora are composed.

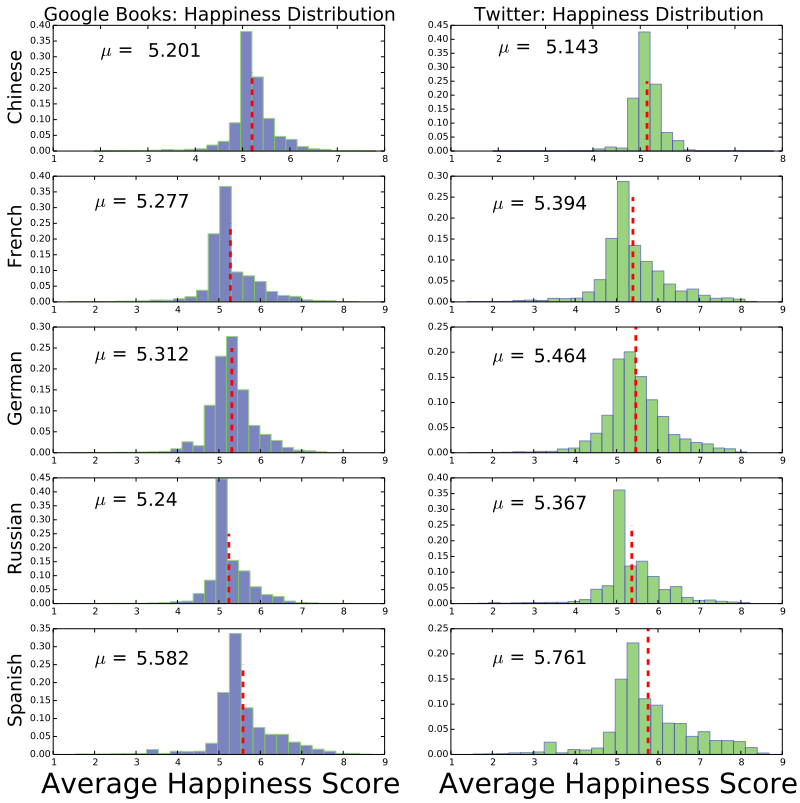
Figure 3.2: Wordlist Kernel Density Estimations



Using frequency distributions from twitter and google books, the density of words are represented as weighted histograms for Spanish, Russian, German, French, and Chinese. The density of positive words ($h_{avg} > 5$) is significantly heavier than negative words. The average of each distribution is slightly above neutral, which is due to the approximately

Zipfian structure of each corpus. Since neutral articles have the highest frequency rank, the densest abundance of words appear in the neutral region, as predicted by by Zipf's law. However, each of these corpora exhibit a clear abundance of positive words over negative words which is a validation of the Pollyanna Hypothesis across language.

Figure 3.3: Google Books and Twitter Happiness Distributions



3.1.2 Multilingual Translation Correlations between Valence

In Figure 3.4, heat map scatterplots illustrate the correlations between evaluations of words translated across language. Each axis represents the happiness score attributed to a word by native speakers from each language. The background color represents whether the row language has an average happiness rating that is greater than the column language. Yellow represents a positive shift, and blue represents a negative shift. The number of words in common, N, the spearman correlation coefficient, R, and the average happiness shift, Δ, are displayed in the upper left corner of each subgraph. The density of points appearing on each heat map are on a color scale from white (densest regions) to black. For example, the row Spanish, column Portuguese yellow subgraph has 3273 translationally stable words, with a Spearman correlation of 0.85, and an average happiness shift of 0.07 which indicates that Spanish survey participants score words as slightly more positive on average than Portuguese participants. Each correlation is significantly positive with p-values < 0.01 , indicating that on average the emotional response of stable words is translationally invariant. Figure 3.5 is has a comparable layout to the previous figure. Here, the histograms of shifted valence scores are given for the ratings of words between each language. These figures help justify the use of translations to help understand the theme conveyed from non-English word-shifts plots.

Figure 3.4: Translation Valence Correlations

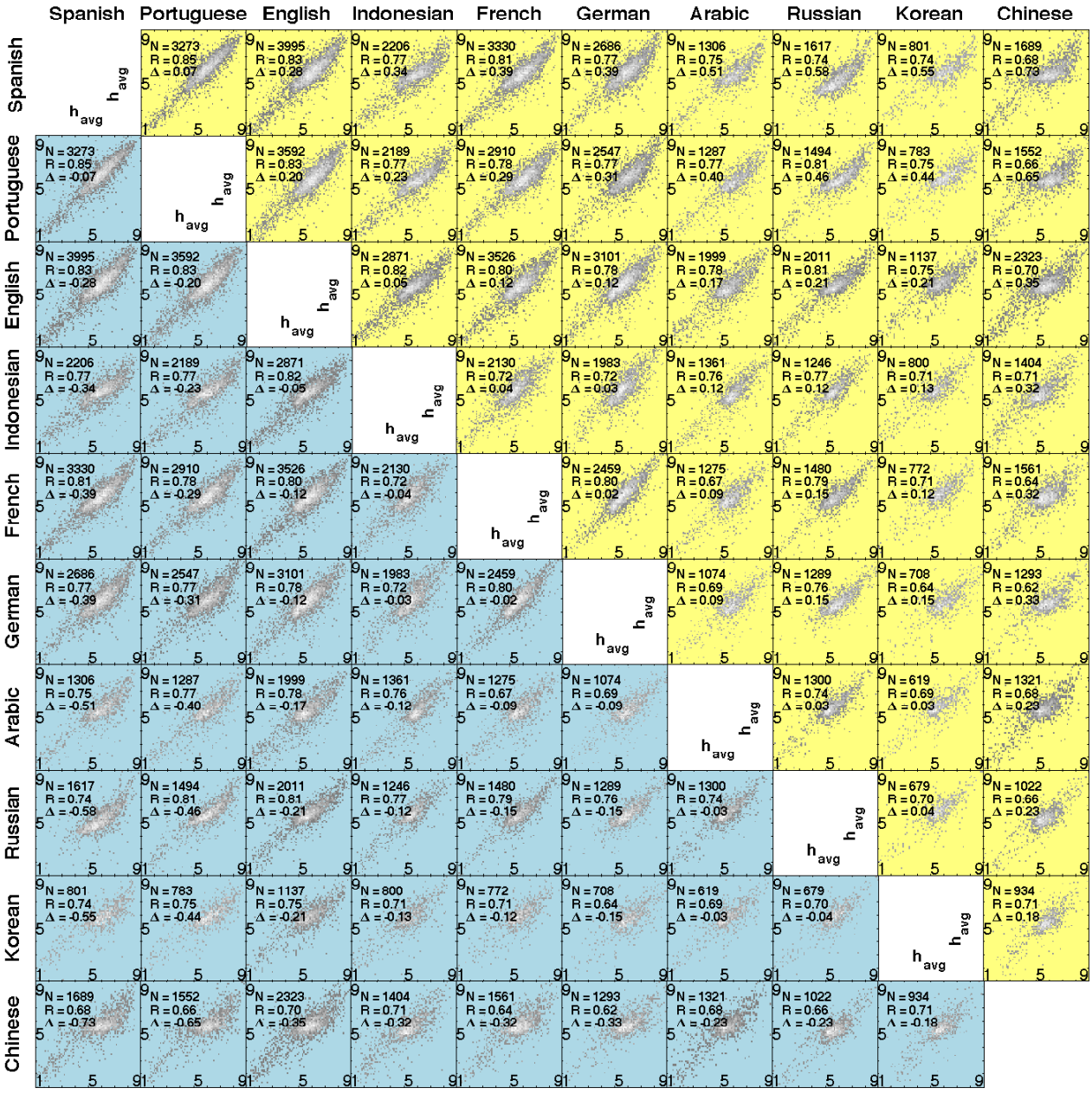
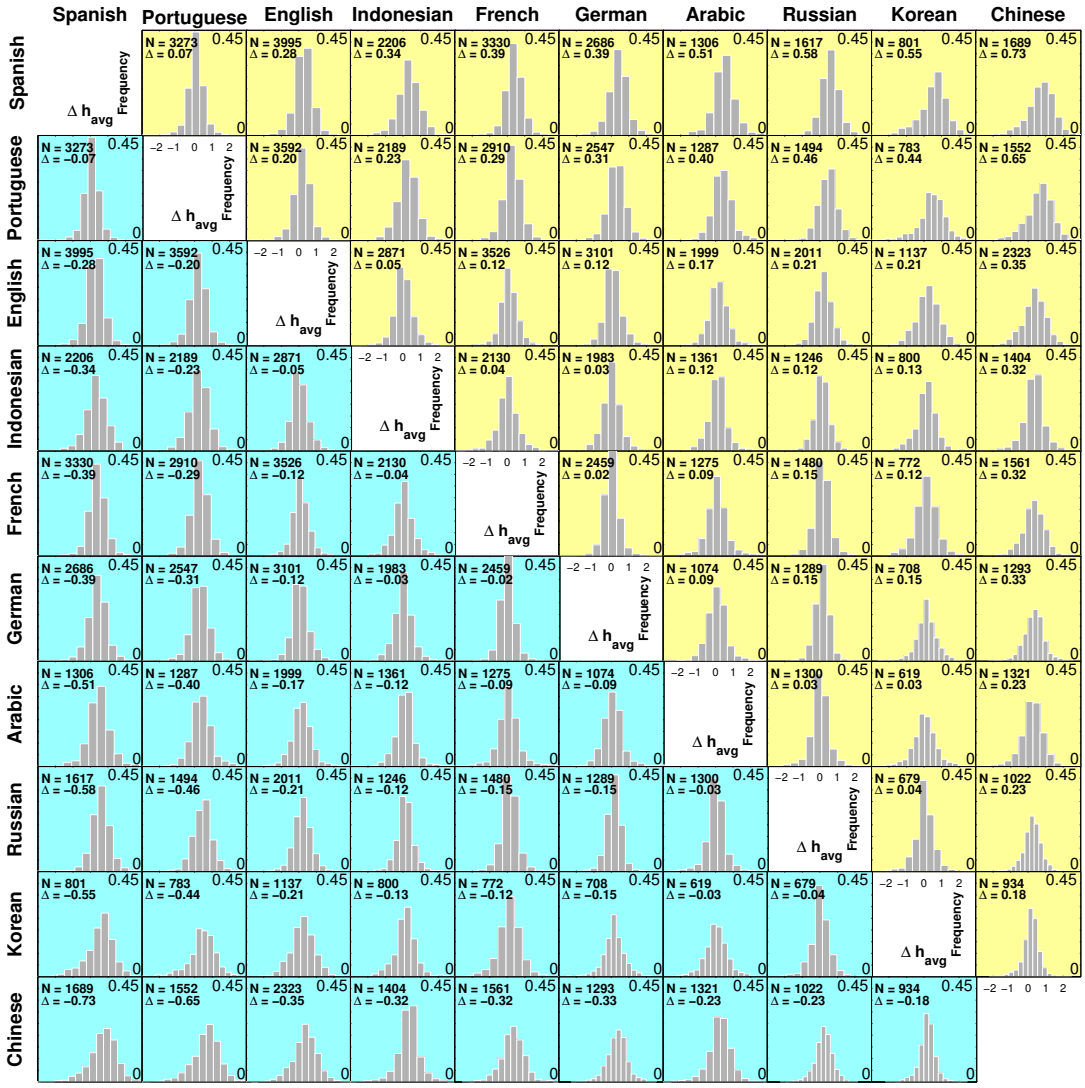


Figure 3.5: Translation Valence Shifts



3.2 20th Century GDP Hedonometric Analysis

GDP per capita is both a measure of the economic well-being of a nation and a indicator of the quality of life for a nation's constituents (Easterlin and Angelescu 2009). (Bentley et al. 2014) has shown this signal can be recovered from the density of emotionally charged words occurring in the literature of a nation. Due to the time difference between the official publication of a novel and the mood of the author during its inception, it is reasonable to expect a time lag between literary emotion and the economic well-being of a nation. Two methods were proposed in (Bentley et al. 2014) that are adapted for this work.

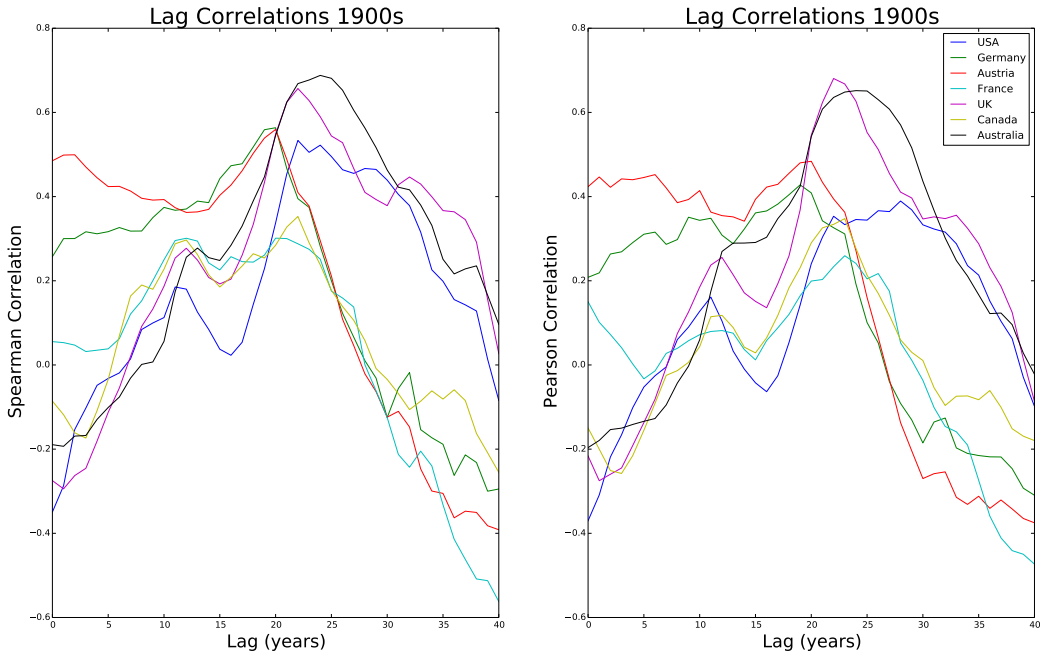
There are some issues to address when restricting to Google Books literature data. The Google Book data set is not separated on a country basis, so correlating GDP data of a specific nation with all literature written in the nation's primary language may obstruct results. Given this restriction, countries with identical primary languages have very similar relationships when correlating with the Google books data set.

3.2.1 Optimal Lag

(Bentley et al. 2014) found an optimal lag of approximately 11 years when correlating economic misery of the US and UK to the average happiness of English literature. The figure below illustrates a similar relationship between the GDP per capita of nation when compared to literature written in the nation's primary language. The USA, Australia, and UK exhibit similar Spearman correlations, with a local optima around 11 years and a global optimum within 25 years. Correlations then steeply decline after this interval. Austria and Germany reach optimal values within a 20 year lag. France and Canada both reach their global optima within 12 years of lag, around the same local optima of the UK, Australia,

and USA. The Pearson correlations plotted to the right exhibit a similar trend. This evidence supports the claim that economic trends are encoded in the mood of literature data. This signal is clearly amplified when introducing a time-lag between the GDP per capita of a nation and the average happiness of literature written in the nation's language.

Figure 3.6: GDP vs Literature Happiness Lag Correlations

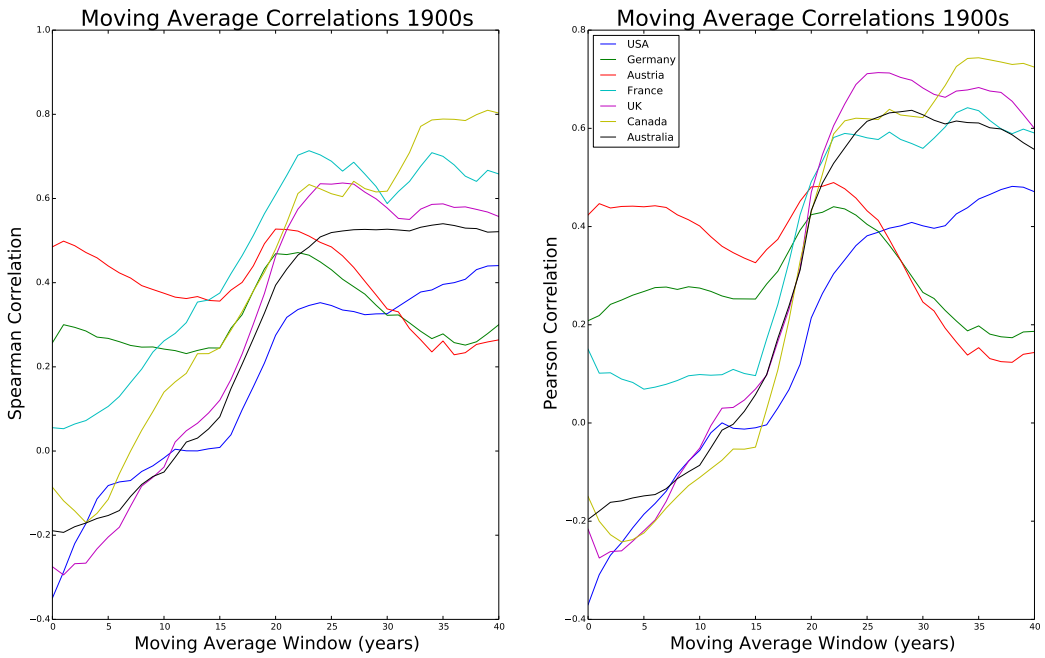


3.2.2 Moving Window Lag

Calculating the moving average for variable time windows on GDP per capita data also yields significant correlations to the literary mood. Introducing a moving average of t years takes each data point occurring at time t_0, calculates the arithmetic mean of the dataset set

from $t_0 - t$ years up to and including the year t_0. In Figure 3.7 the moving average window is varied by year and the corresponding correlation coefficient between the averaged GDP data set and the average happiness of literature data are plotted. Correlations have significantly increased in comparison to introducing a discrete time lag. A similar result is presented in (Bentley et al. 2014) when correlating the economic misery of nation with the literary mood for the US and Germany. Here, however, the correlations increase with the moving average window until reaching an optimal window size, which seems to be comparable to the global optima of Figure 3.6.

Figure 3.7: GDP vs Literature Happiness Moving Average Correlations



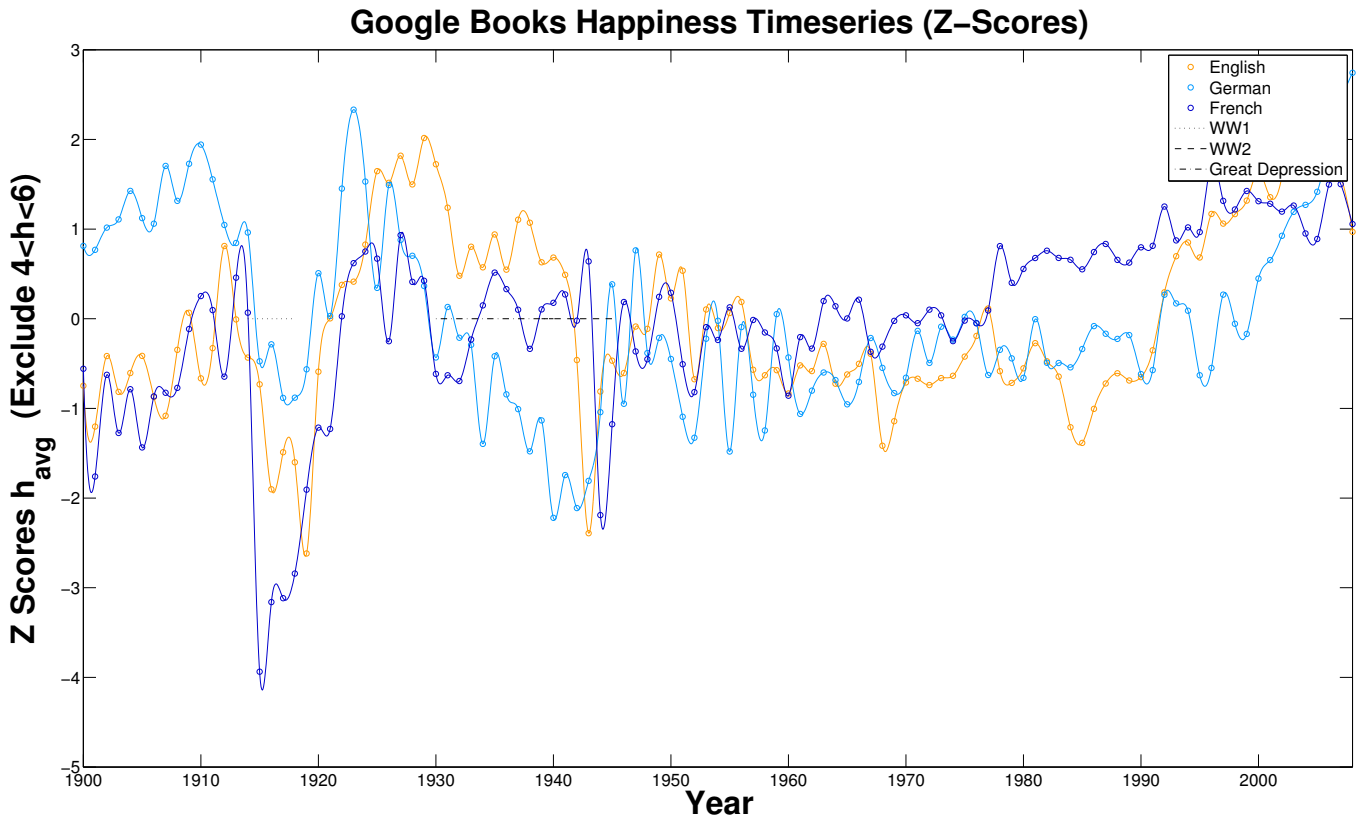
3.3 Hedono-History Tour

Time-series plots of the average happiness value per year of 20th century literature yield insight to the time-period surrounding notable historical events. Time series data alone can only lead to speculation of how an event is impacting the literary mood. The prominent words that are inducing the average happiness shift illustrate the historical events that are being reported through literature. Although the Google Books data set spans from the 1500's through the 20th century, the densest region of lexical data occurs over the 20th century. Since hedonometrics yield more substantial results when averaging over a particularly large sample of text, (Dodds and Danforth 2009) , only the 20th century is included in this analysis.

3.3.1 Time-series Analysis

In Figure 3.8 the Z-scores of the average happiness value per year are plotted for English, German, French, Russian, and Spanish literature. As in (Acerbi et al. 2013) using the Z-scores of each data point help identify grave shifts from the global mean of the data set. This method identifies historical events as outliers on the time series graph. Each data point is connected via cubic spline interpolation techniques.

Figure 3.8: Google Books Happiness Time-series



Two universally negative events occur during each of the World Wars, which are indicated on the time series. The Great Depression also had a severe negative impact on German and English literature. The depression was not as severe in France, however some signal of this economic event is still prevalent. Due to the vast number of books encompassing this dataset coupled with annually reported time-stamps, only large scale historical events can be recovered from the happiness time-series. Using word shift graphs, the nature of the event that is prevalently discussed through literature becomes obvious. The following sections use word shift graphs to identify the key words that influenced the emotional decline in literature over the World Wars and the Great Depression.

3.3.2 World War I

Word-shift graphs surrounding the time frame of each World War reveal which specific words have the most significant contribution to the negative shift in literary emotion. To demonstrate the power of word shift graphs to visualize these historical events, the analysis is focused on France, Germany, and English literature as these countries were major players in the world war. Two different techniques are implemented to create word shift graphs. The first uses reference and comparison periods that are compiled from a full decade of literature data. Using a full decade accounts for the lag present between publication and documentation of the actual historical event. The second technique uses the full 20th century dataset as the reference period and a single decade for comparison. The reference period must be delicately chosen in order to obtain an optimal representation of the words that have the most significant impact during the time period of interest.

Figure 3.9: English and German Word-shifts 1900s versus 1920s

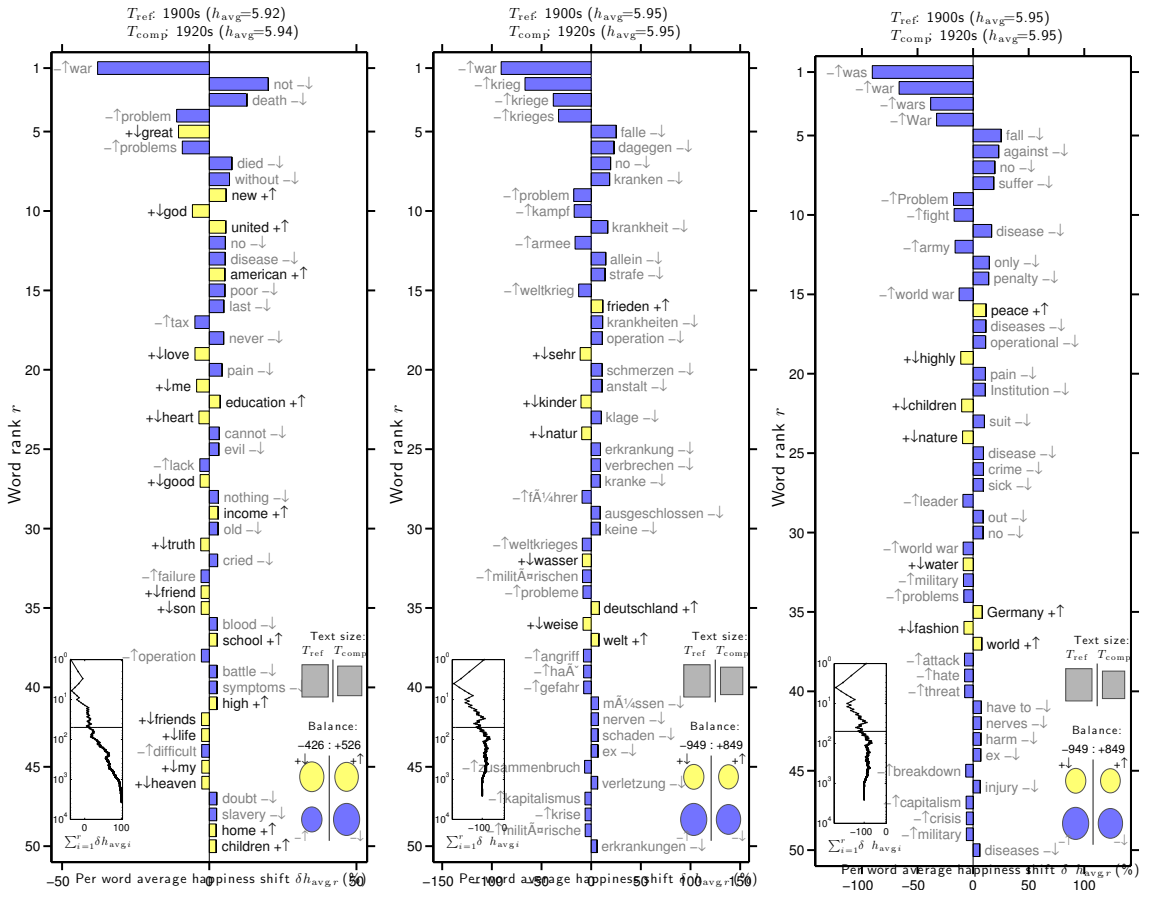


Figure 3.9 compares word shifts in English and German literature surrounding World War I (1914-1918). The reference period comprises all literature between 1900 and 1910 while the comparison period comprises 1920-1930. The graph on the far left is composed from English literature and the remaining graphs are obtained from German. The graph on the far right is a translation of the central graph from German to English for reference. Each word shift graph reports 'war' as the greatest contributor to the negative emotional shift during this time-period. There are several other indicators on the German word shift

plots that are recovering the signal from the world war including an increase in the density of the words translating to 'war', 'problem', 'fight', 'army', 'world-war', and 'military'.

Figure 3.10: English and German Word-shifts 1910s versus 1920s

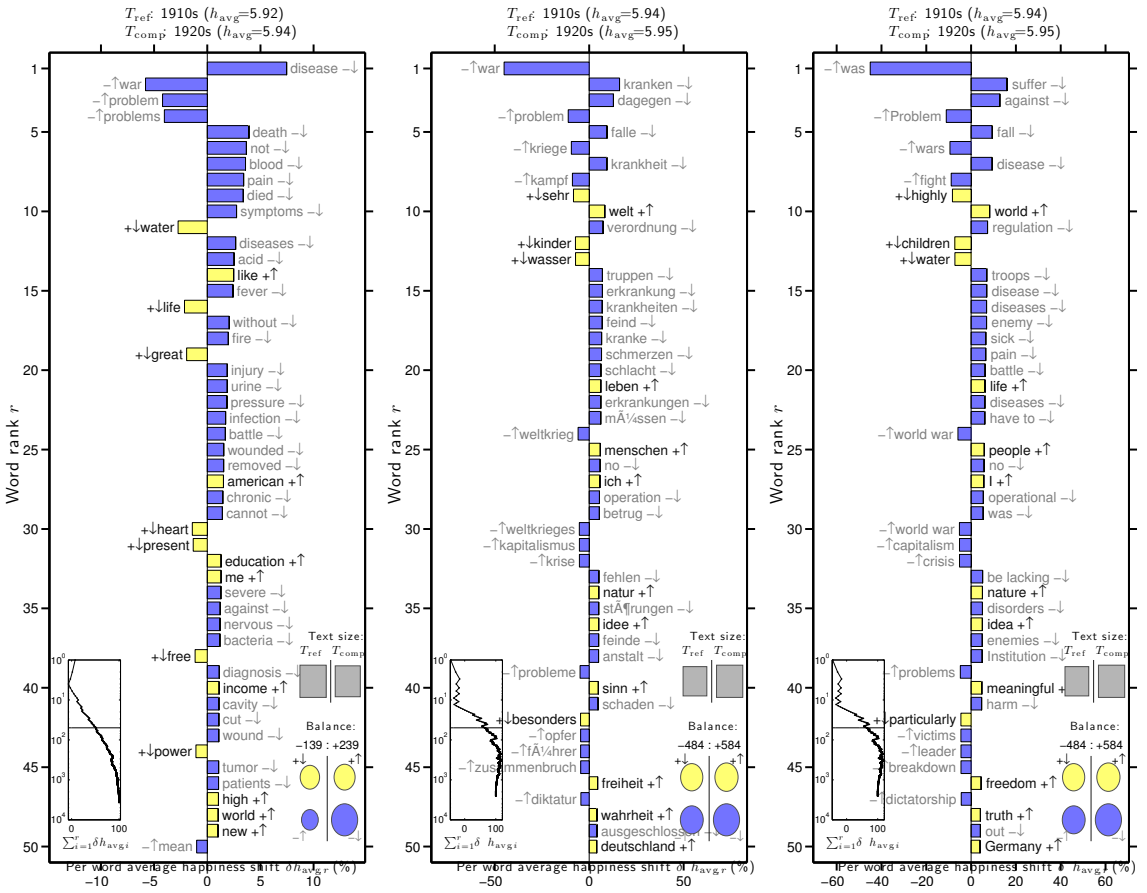


Figure 3.10 uses the World War decade as reference for comparison to the decade directly following the war. English literature (far left) shows an increased frequency of the words 'war' and 'problem' however a decreased frequency of the 'words', 'disease', 'death', 'blood', 'pain', and 'died' which indicate that the world war was no longer the most prominent topic in literature. This indicates that the decade surrounding World War

I had a high density of aggressively charged negative words: 'battle', 'blood', 'pain', 'disease', and 'wounded'. The German word shifts (center and far right) portray the same theme. There is a significant decrease in the density of words that translate to 'suffer', 'disease', 'troops', 'enemy', 'pain', and 'battle' in the decade following the World War.

Figure 3.11: French Wordshifts 20th century versus 1910-1920

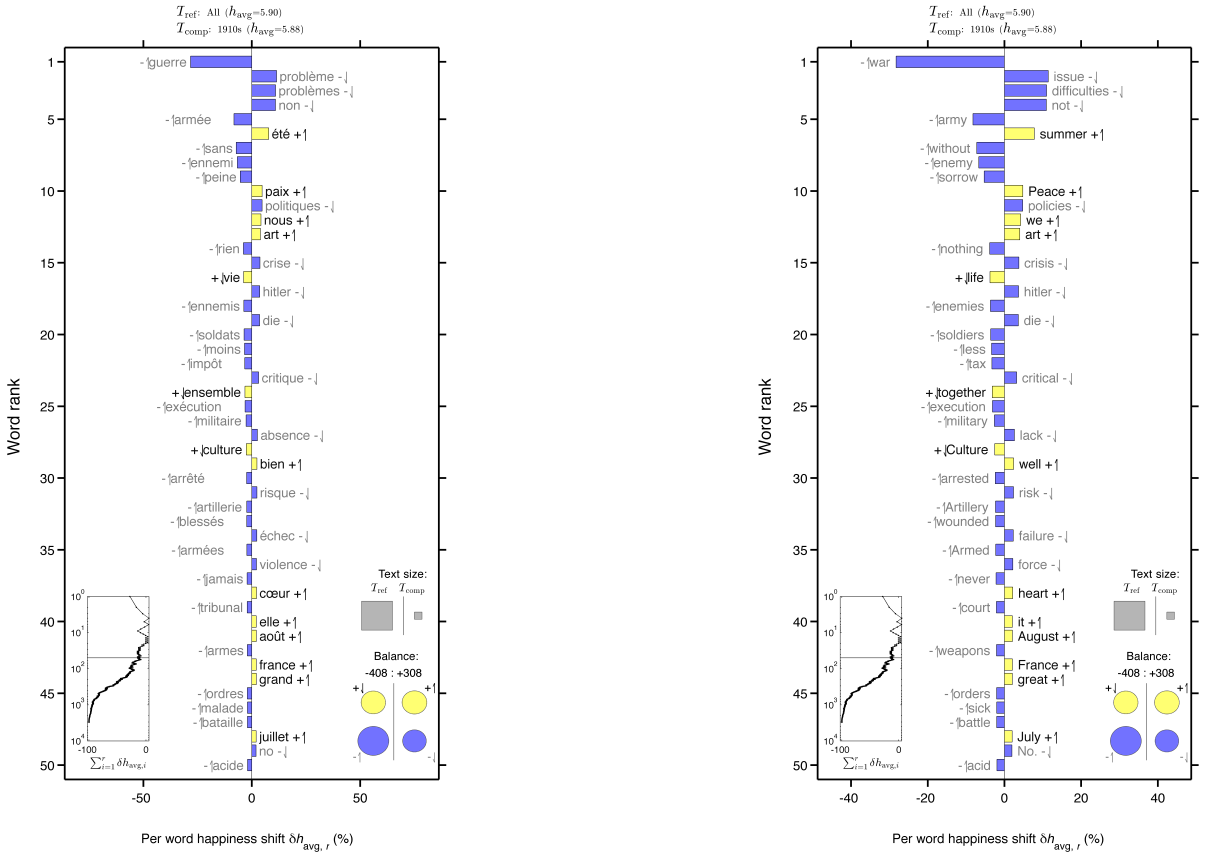


Figure 3.11 are word shifts created from the French literature corpus. As per the previous figures, the French words causing the shift in emotion appear on the left and their direct English translations appear on the right. Here, the reference period is taken as the

entire 20th century literature dataset and the comparison period ranges from 1910 to 1920. Again we see an increase in aggressively charged negative words: 'war', 'army', 'enemy', 'sorrow', 'soldiers', 'artillery', 'wound', 'battle', 'execution', 'acid'. Notice the decrease (or lack of existence) of the negative word 'hitler' which is prevalent in future decades surrounding World War II. This is more evidence that the negative emotional shift in the time-series during this period is directly related to World War I.

3.3.3 Great Depression: 1930-1940

The economic collapse of the 1930's, known as the Great Depression, impacted nations across the world, (Smiley). A high density of literature from this era used an abundance of negatively charged words as shown on the time series analysis from the previous section. Word-shift graphs centered around this era demonstrate the most significant contributors of this decline in literary happiness.

Figure 3.12: American/UK Great Depression

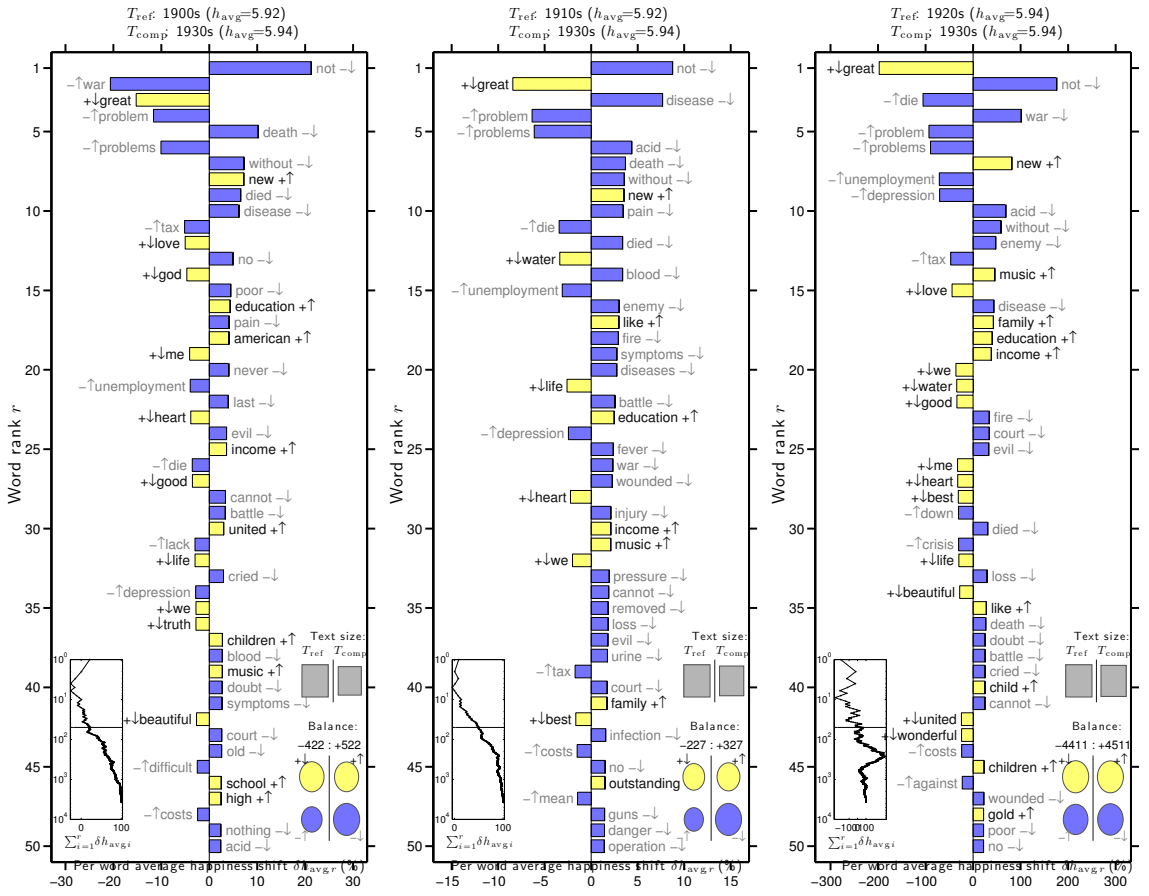
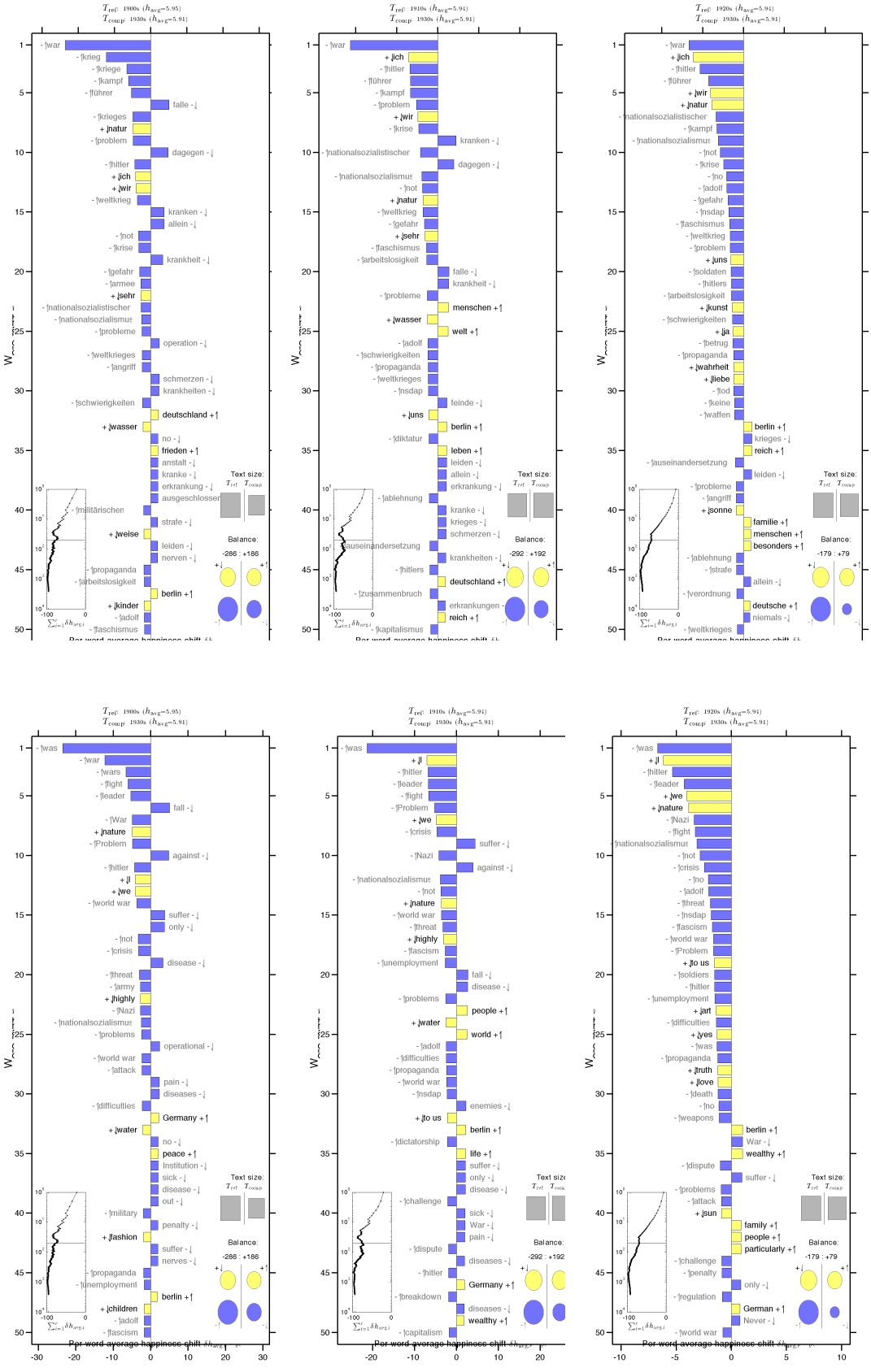


Figure 3.12 displays several different reference periods against the 1930s for comparison. Since World War I occurred after 1910, an increase in the use of the word 'war' has the highest contribution to the negative shift. Combat jargon ('army', 'battle', enemy') decrease in the next two graphs, since by this point the World War had significant documentation. Notice across each graph the increase of 'problem', 'tax', 'unemployment', 'depression', 'costs' which is indicative of a significant economic event.

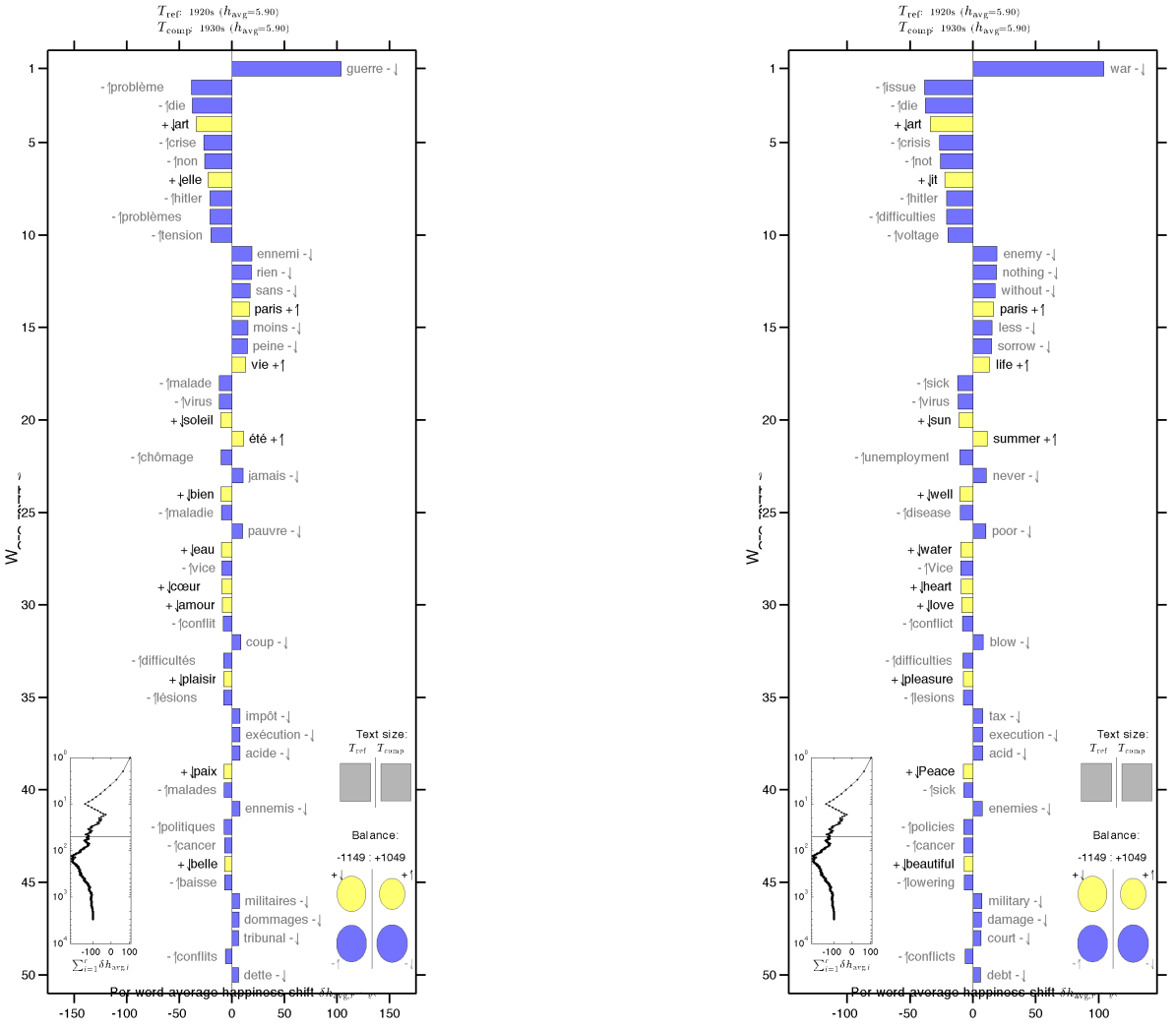
Figure 3.13: German Great Depression Word-shifts



The Great depression had a grave impact German economics which is reflected by the literature of the time period. The top row of Figure 3.13 are the German word shifts and below each graph are the English translations of each of these German words. The reference and comparison periods are chosen in the same manner as per the English word shifts. Here there is a noticeable increase in 'unemployment', 'problems', 'difficulties', 'regulation', and 'crisis' indicative of economic literature. At the peak of the German depression, unemployment rates rose to upwards of 30%, (Smiley). This economic weakness pushed Germans to rally for a regime change, which paved the way for Hitler and the fascist Nazi party to attain power, (USA). The rise in 'Nazi', 'adolf', 'Hitler', 'leader', and 'facism', is a reflection of this event.

The French depression was much less dramatic, however word shifts plots of the 1930s show some reflection of economic stress. Notice the rise in 'issue', 'crisis', 'difficulties', 'sick', and 'policies'. The dramatic increase in the word 'crisis' may be from literature reporting the "6 February 1934" crisis that incited a regime change, (Millington 2014). This crisis was a direct result of economic hardships. Although the signal of the depression is fainter in French literature, it can still be identified with word-shifts from the time period.

Figure 3.14: French Depression



3.3.4 World War II : 1939-1945

World War II spanned from 1939 to 1945. This particularly dark historical event had a significantly negative impact on literary mood as illustrated in time-series of Figure 3.8. Word shift plots indicate that the World War was the culprit of the abundance of negatively emotional literature. In Figure 3.15, several reference periods are chosen for comparison of the 1940s. 'War', 'die', 'problem', 'enemy', 'bomb' are reasonable indications of the war's documentation over English literature.

Figure 3.15: English World War II Wordshifts

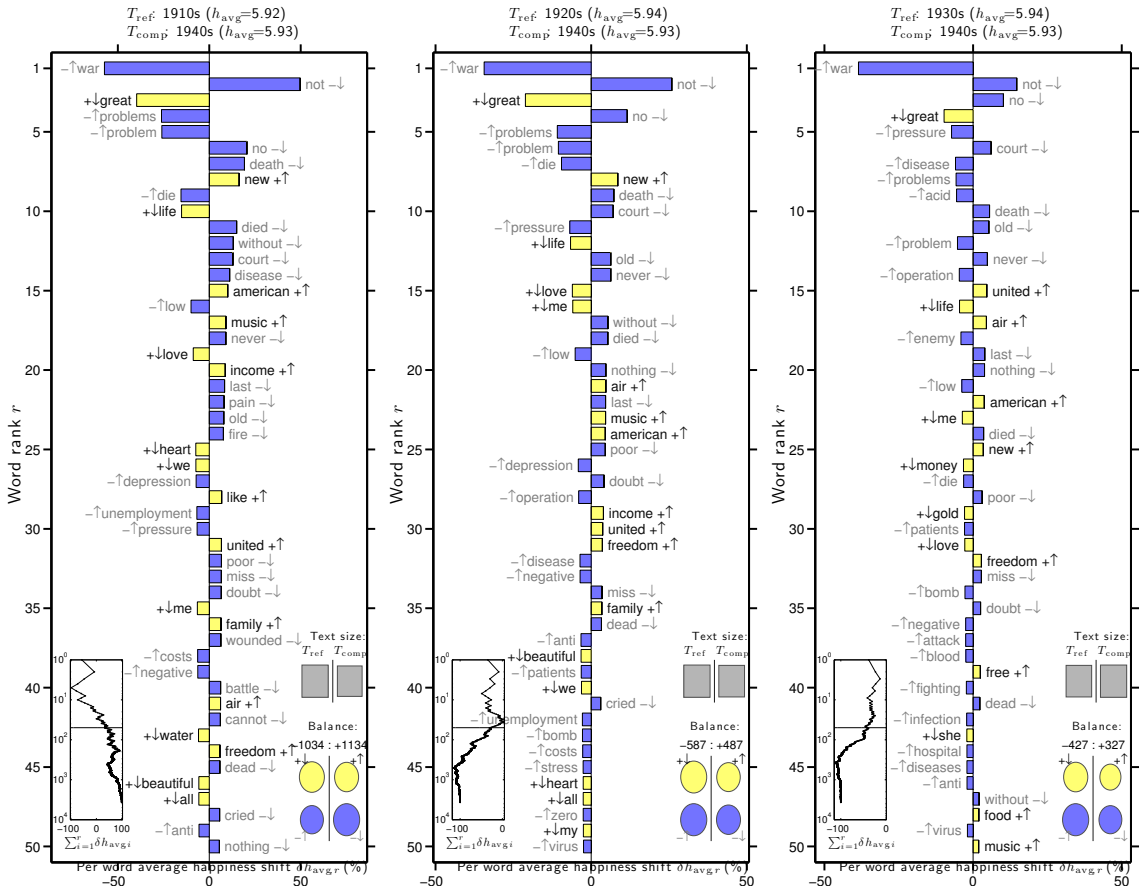


Figure 3.16: German World War II Wordshifts

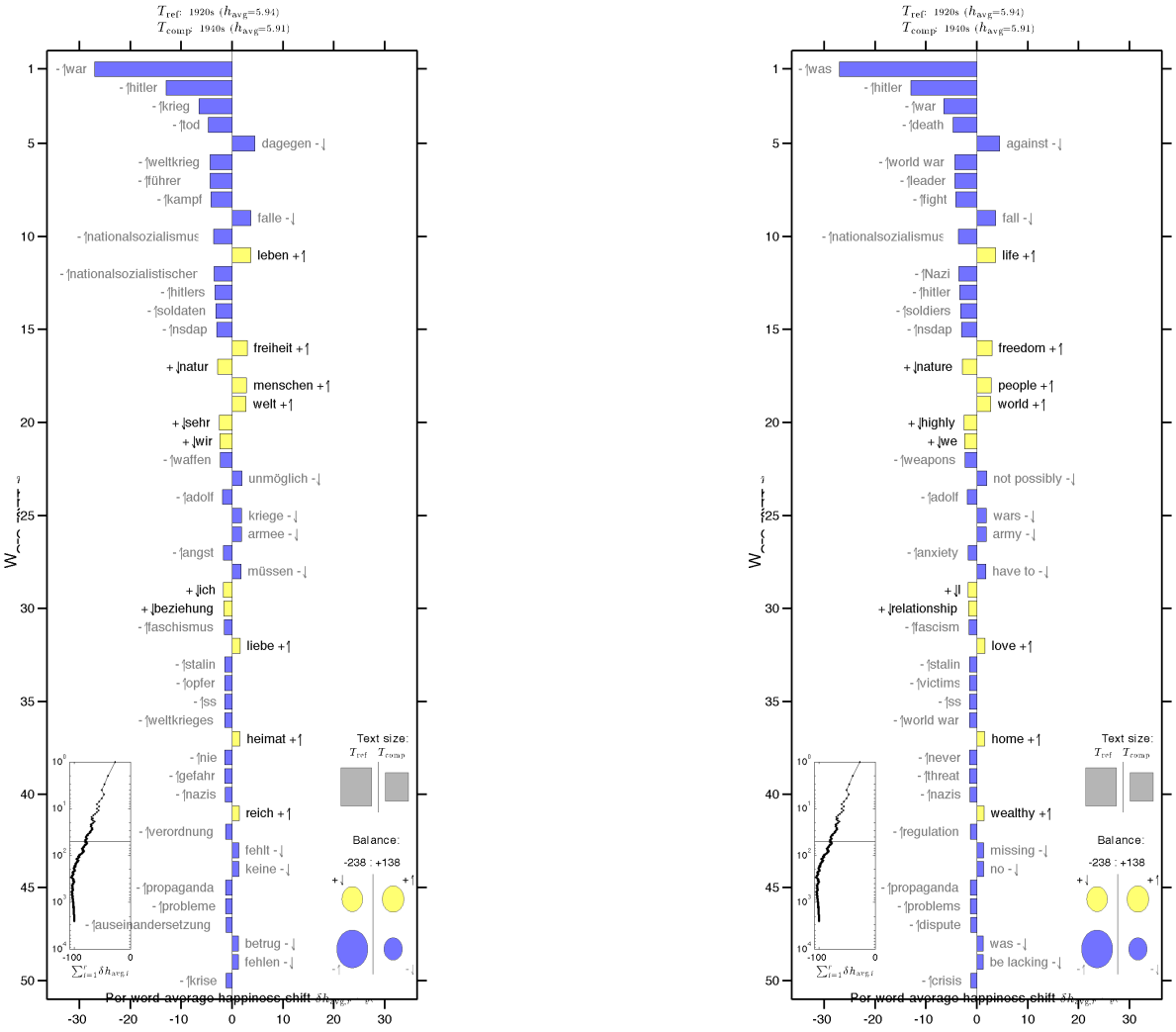
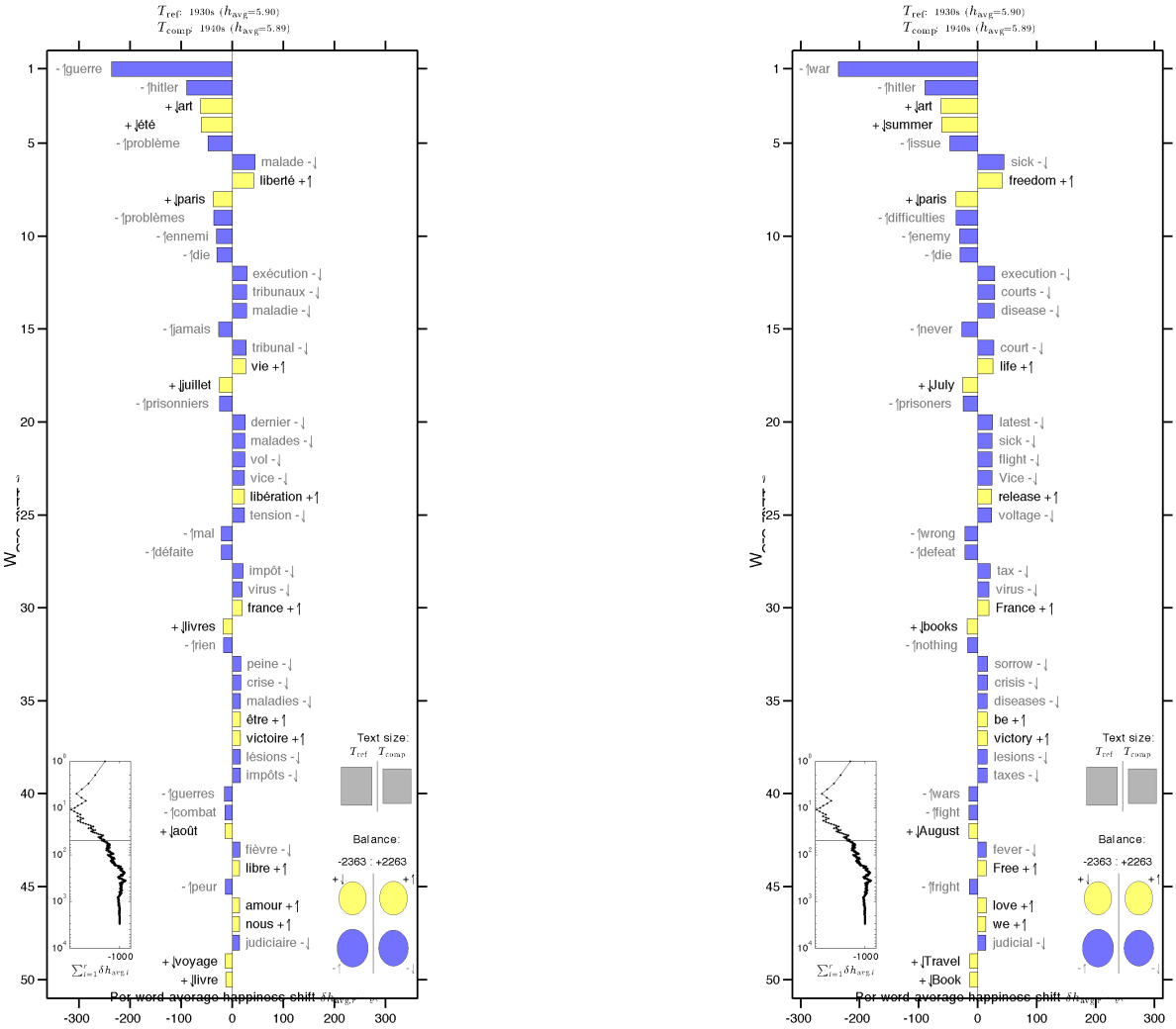


Figure 3.16 presents a German word shift plot from this era. Here, the 1920s are used as a reference period for the 1940s. Translations report an increase of 'world war', 'hitler',

'fight', 'Nazi', 'weapons' ,'adolf', 'SS', and several other combative words that are specifically related to World War II.

Figure 3.17: French World War II Wordshifts



French word shifts (Figure 3.17) display the same theme. Here the 1920s are used in reference to compare the 1940s. Translations indicate an increase in 'hitler', 'war', 'enemy','die', 'difficulties', 'defeat' and many other indications of the war's documentation. Each of these word shifts actively convey the literary theme to match the trends in the Happiness time series graphs. Understanding the emotionally charged words of each of these time-periods is an efficient way to understand the most significant contributors in swinging the emotional literary mood.

3.4 2012 Summer Olympics

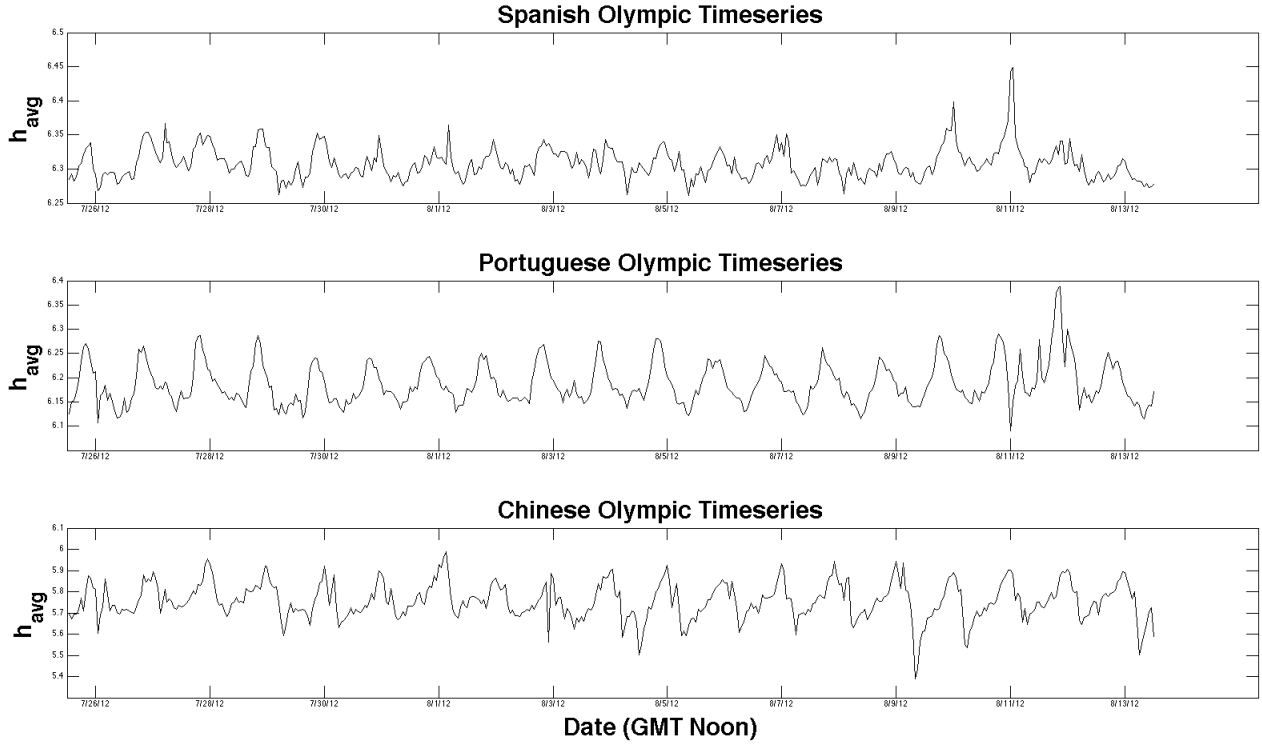
The hedonometer has the capability to identify large scale historical events by targeting emotional outlier years in literature. We now shift our focus to identifying time-sensitive events that have been reported by users of social media. Specifically, we investigate the activity of Spanish and Portuguese tweets that occurred during the 2012 Summer Olympics, which occured between July 26th and August 13th. This event was a prominent topic on twitter. Due to the advent of smartphones, tweets report events that are happening in real time. This time period is investigated with a hedonometric analysis consisting of happiness time series and word shifts plots.

3.4.1 Olympics Happiness Time-series

Spanish, Portuguese, and Chinese tweets are collected for each hour spanning 7/26/12 through 8/13/12. Using the hedonometer, average happiness is calculated as function of the hour of each tweet's composition. Figure 3.18 presents the happiness time series of tweets during the Olympic games. Word shift graphs convey that highly emotional tweets

discussing the outcome of each olympic award are likely responsible for many of these local optima.

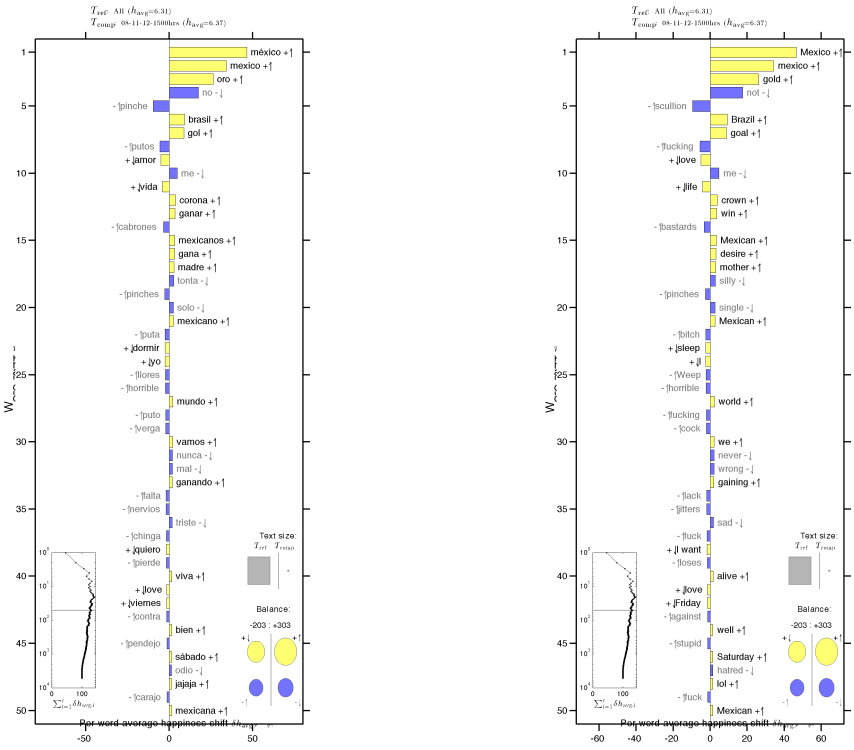
Figure 3.18: Olympic Happiness Timeseries



3.4.2 Spanish Wordshift Plots

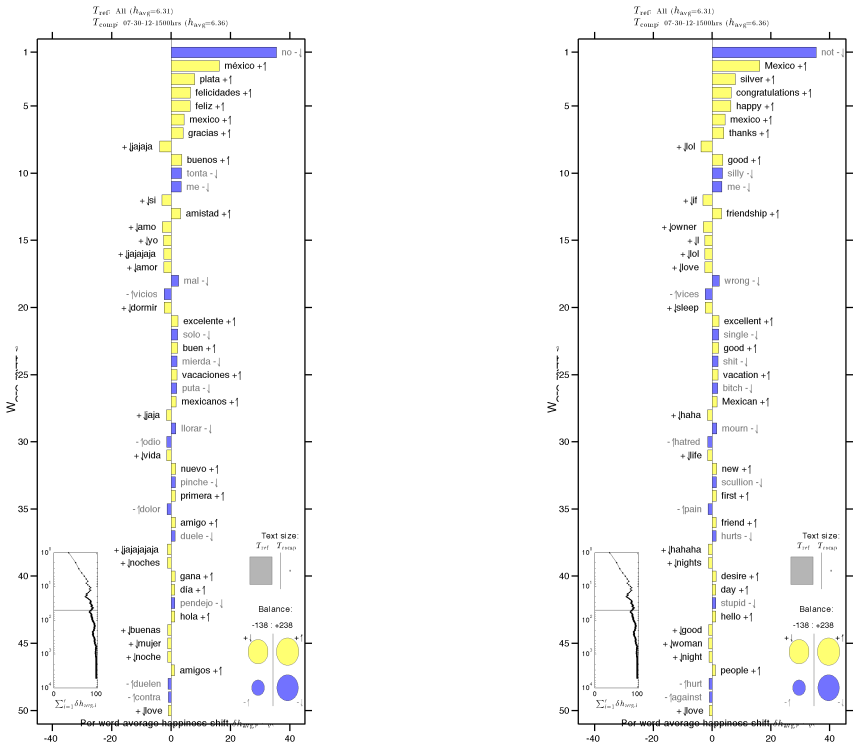
In each of the following word shifts, the full Olympic data set is used as a reference period to analyze the hour of the olympic award. The most notable outlier on the time series plot occurs during 8/11/12 when Mexico won the gold medal in soccer. The words responsible for the shift include 'Mexico','gold', 'goal'. This shows the outcome of this olympic game had a dense representation twitter, and was highly responsible for the sharp incline in valence.

Figure 3.19: 8-11-12: Mexico- Soccer Gold Medal



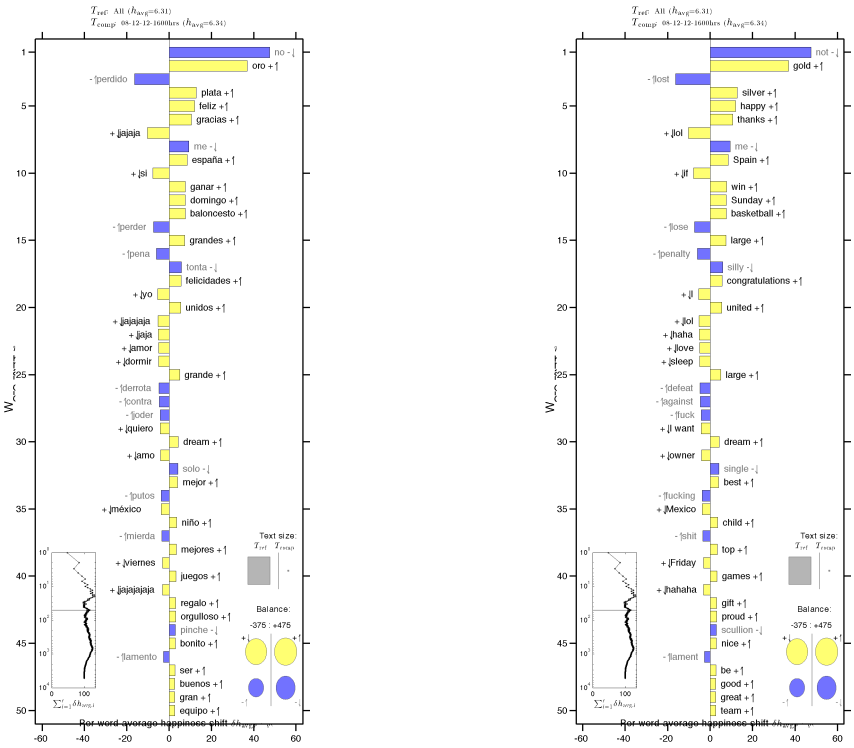
The most notable outlier on the time series plot occurs during 8/11/12 when Mexico won the gold medal in soccer. The words responsible for the shift include 'Mexico','gold', 'goal', as shown in Figure 3.19.

Figure 3.20: 7-30-12: Mexico- Diving Silver Medal



Another notable optima occurs at 7/30/12 precisely when Mexico earned the silver medal in diving. An increase in 'silver', 'Mexico', 'congratulations', 'happy', and 'excellent' are strong indications that this award also had a significant positive impact on tweets.

Figure 3.21: 8-12-12: Spain- Basketball Silver Medal



The wordshifts in Figure 3.21 illustrate the response to Spain earning the silver medal in basketball. Here the words 'silver',' Spain','win','games', 'team', 'penalty' , and 'lost' indicate a high density of tweets were describing the olympic event.

Figure 3.22: 8-11-12: Brazil Volleyball Gold and Soccer Silver Medals

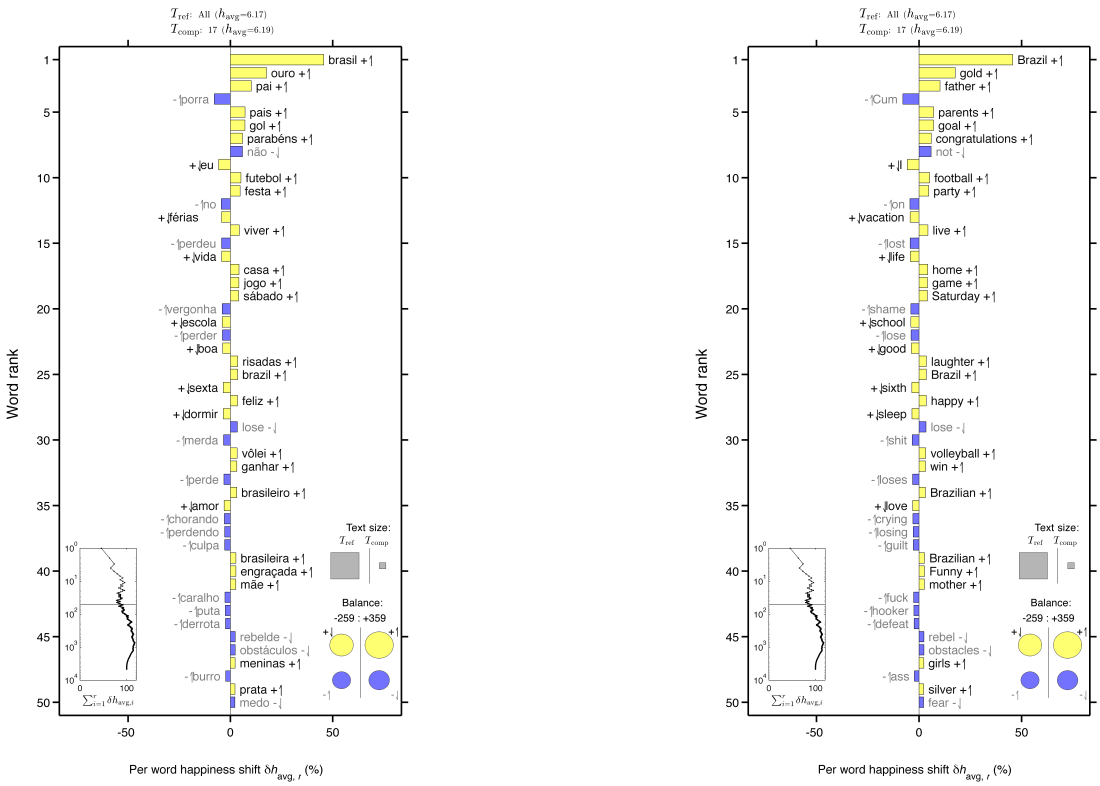


Figure 3.22 presents the word-shift of Portuguese Tweets during the 17th day of the Olympics is used for comparison against the full Olympic data set. On this day, Brazil won a gold medal in women's volleyball and a silver medal in soccer. Translations of the words appearing in Portuguese indicate an abundance of 'Brazil', 'gold', 'silver', 'goal', 'congratulations', 'volleyball', and 'football'. These words serve as an indication that the emotional shift in the time series graph is indeed due to Brazil's performance in the Olympic Games.

Chapter 4

Conclusion

We have demonstrated the power of this newly adapted multi-lingual hedonometer. Wordlist valence distributions across language coupled with the proposed time series analysis and wordshift graph methods can uncover the underlying story of large databases of mixed written text. A multi-lingual positivity bias across literature and the twittersphere has been verified for several languages. Using these valence distributions, economic trends have been correlated with the emotional mood of literature.

The valence distributions of these new wordlists have validated the Pollyanna Hypothesis across literature from the Google Books corpus and twitter. The abundance of positive words in literature and tweets are indications of the natural human tendency to express positive events in daily life. The subset of translationally stable words tend to preserve their emotional charge between languages. Hence, the stable words on each wordlist have an above average tendency towards translation invariance. This result is another validation of lexical preservation of meaning across language.

Since the corpora composing each wordlist distribution were approximately Zipfian, each of these multi-lingual word distributions can be effectively implemented analogously

to the LabMT word set. Ignoring stop words allows for informative time-series plots that can indicate significant historical events from the shift in abundantly emotional literary text. These new wordlist distribution also found evidence that economic trends are encoded in the multi-lingual literary mood. A natural lag between publication and literary mood indicates that the economic well being of a nation and its literature are connected. When restricted to the twitter sphere, the signal of small scale events can be recovered at the hourly resolution. Using multi-lingual word shift graphs historical events large and small can be identified with outliers in the happiness time-series of several different corpora. Future work can incorporate the geographic locations of tweets in each of these languages to extend the work of (Frank et al. 2013, Mitchell et al. 2013). These new multi-lingual wordlist have the potential to explore a vast number of interesting computational linguistic topics.

References

Acerbi, A., V. Lampos, and R. A. Bently (2013). Robustness of emotion extraction from 20th century english books. *Big Data, 2013 IEEE International Conference*.

Amazon-Mechanical-Turk. Amazon's mechanical turk service. Available at https://www.mturk.com.

Appen-Butler-Hill. Global Network of Linguists. available at http://www.appen.com/.

Bentley, A., A. Acerbi, P. Ormerod, and V. Lampos (2014). Books average previous decade of economic misery. *PLOSone*.

Boucher, J. and C. E. Osgood (1969). The pollyanna hypothesis. *Elsevier 8*(1), 1–8.

Bradley, M. and P. Lang (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*.

Brants, T. and A. Franz (2006). Web 1t 5-gram version. *Linguistic Data Consortium, Philadelphia Linguistic Data Consortium, Philadelphia Linguistic Data Consortium, Philadelphia*.

Dodds, P. and C. Danforth (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Springerlink.com*.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOSone*.

Easterlin, R. (1974). *Does economic growth improve the human lot? Some empirical evidence.*, Volume Nations and households in economic growth: Essays in honour of Moses Abramowitz, pp. 89–125. New York: Academic Press.

Easterlin, R. A. and L. Angelescu (2009). Modern economic growth and quality of life: Cross sectional and time series evidence. Technical report, Mimeo, University of Southern California.

Frank, M. R., L. Mitchell, P. S. Dodds, and C. M. Danforth (2013). Happiness and the patterns of life: A sutdy of geolocated tweets. *physics.soc-ph*.

Google-Labs. Google labs ngram viewer. available at http://ngrams.googlelabs.com.

Google-Translate. available at translate.google.com.

Kloumann, I., C. Danforth, K. D. Harris, C. Bliss, and P. Dodds (2012). Positivity of the english language. *PLOSone*.

Lang, B. M. (1999). Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. *Technical report c-1, University of Florida, Gainesville, FL.*

REFERENCES

Lenhart, A. and S. Fox (2006). Bloggers: A portrait of the internet's new storytellers. *Technical report. Pew Internet and American Life Project*.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, and M. K. G. andWilliam Brockman (2010). Quantitative analysis of culture using millions of digitized books. *Science*.

Millington, C. (2014). 80 years ago today: 6 february 1934, french fascists topple government. http://frenchhistoryonline.com/2014/02/06/80-years-ago-today-6-february-1934-french-fascists-topple-government. 2014.

Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*.

New-York-Times. Olympics 2012. http://london2012.nytimes.com/.

Petersen, A. M., J. Tenenbaum, S. Haavlin, H. E. Stanley, and M. Perc (July, 9 2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*.

Redondo, J., I. Fraga, I. Padron, and M. Comesana (2007). The spanish adaptation of anew (affective norms for english words). *Behavior Research Methods 39*, 600–605.

Smiley, G. The concise encyclopedia of economics: Great depression. www.econlib.org/library/enc/greatdepression.html. 2014.

Soares, A. P., M. Comesana, A. P. Pinheiro, A. Simoes, and C. S. Frade (2012). The adaptation of the affective norms for english words (anew) for european portuguese. *Behavioral Research Methods 44*, 256–269.

The-World-Bank. Gdp per capita. available at http://data.worldbank.org/.

Twitter. Twitter api. available at http://dev.twitter.com.

USA. United-states-holocaust-memorial-museum. hitler comes to power. http://www.ushmm.org/outreach/en/article.php?moduleid=10007671. 2014.

Zipf, G. (1949). *Human Behaviour and the Principle of Least-Effort*. MA: Addison Wesley.