Measuring Subjectivity in History Textbooks

Onyi Lam[1] and Eddie Lin[2]

[1]Pew Research Center
[2]Center for Data Science and Public Policy, University of Chicago

May 31, 2017

Abstract

History textbooks provide a lens through which students view the nation's past. Government, especially that of authoritarian regime, has an incentive to present historical content in a more subjective narrative to influence students' political views. This paper tests this claim by considering the problem of measuring subjectivity history textbooks in China, Hong Kong and Taiwan. By doing so, it illustrates a linguistic manifest of persuasive communication that has been less studied in the literature: text subjectivity. Using insights developed in the field of sentiment analysis, we find empirical evidence that history textbooks in mainland China exhibit stronger degree of subjectivity than history textbooks used in Hong Kong and Taiwan. Specifically, the paper measures subjectivity by calculating the adjective ratio, the ratio of positive to negative phrases, ratio of Chinese Communist Party (CCP) to Kuomintang (KMT) entities mentions and word embedding method that measures distance from phrases of interest such as the Chinese Communist Party to other adjectives.

1 Introduction

History textbooks provide a lens through which students view their nation's past. In absence of political influences, history textbooks should provide an unbiased, objective and accurate account of the nation's past events. In reality, however, governments can use history textbooks "as ideological tools to promote a certain belief system and legitimize an established political and social order." (Apple and Christian-Smith (1991, 10)). Authoritarian regimes, in particular, have a strong incentive to homogenize political views of their citizens so that they do not wish to rebel against the regime. As such, content in history textbooks in authoritarian regimes is especially vulnerable to government manipulation.

Compared to the mass media, history textbooks are arguably more formative in shaping one's support to the regime and easier to manipulate. There are several reasons. Unlike mass media,

[1]Yolanda Xue provides excellent research assistance.

history textbooks are mandatory for students. Students are also incentivised to learn the material well in order to do well in exams. Furthermore, history textbooks are studied at an age when values are most influenced[1]. These important characteristics of history textbooks mean that understanding how history is being portrayed can help us understand the divergence in political views between citizens residing in different regions with shared history.

This paper measures the degree of subjectivity in history textbooks by using insights and tools developed in the field of natural language processing. Doing so also allows us to illustrates one linguistic channel through which persuasive communication manifests. Subjectivity in natural language processing refers to aspects of language used to express opinions and evaluations (Banfield 1982; Wiebe 1994). Subjective remarks can come in a variety of forms, including opinions, rants, allegations, accusations, suspicions, and speculation (Riloff et al, 2013). Though subjective remarks and bias can have similar meanings in some contexts, in this paper, we consider subjective remarks as text that emphasizes the writer's views and feelings on a given subject and bias as factual but incomplete account of that subject. In other words, the main difference between bias and subjective remark is that subjective text contains less factual information such as numbers and details of the state. However, it should be noted that objective text, in absence of the author's own opinions and evaluations, can be completely biased.

We study high school history textbooks used in mainland China, Hong Kong and Taiwan. China observers have noted that Beijing uses "history education as an instrument for the glorification of the party, for the consolidation of the PRC's national identity, and for the justification of the political system of the CCP's one party rule." (Wang, 2008) Comparing textbooks between mainland, Hong Kong and Taiwan allows us to study and contrast the attempt of governments in these three regions to influence the citizens' view on the mainland Chinese regime. To put the degree of subjectivity in perspective, we also include in our quantitative analysis a translated copy of "Search of Modern China" (SMC) - a book on Chinese history written by the authoritative figure in China studies, British-American scholar Jonathan Spence. SMC is frequently assigned as a reference in Chinese history courses in universities in the United States. In addition to the modern history textbooks, we also analyzed two textbooks prior to major textbook reforms in the mainland and Taiwan to study the change in subjectivity over time.

The interest in history textbook is not new. In particular, history textbook in East Asia has been a source of international tension and domestic controversies. Previous studies have noted the divergence of historical account of world war II in textbooks used in East Asia (Sneider, 2013). Despite Beijing and Seoul's criticism on Japanese nationalist effort in whitewashing

[1]. Studies have suggested that political views are influenced most in early years. The most relevant theory in this respect, the impressionable years hypothesis, states that core attitudes, beliefs, and values crystallize during a period of great mental plasticity in early adulthood (the so-called impressionable years) and remain largely unaltered thereafter.

Japan's atrocities in world war II in some controversial textbooks, mainstream Japanese textbooks are found to be relatively devoid of overt attempts to promote patriotism. In contrary, official mainland Chinese textbooks are found to promote a nationalist view of the past. Controversy over history textbooks also took place in Taiwan where critics of a proposed textbook reform said the changes "China-centric" and deny the island its own perspective.

While there exists several studies that measure bias in mass media (Gentzkow and Shapiro, 2010; Groseclose and Milyo, 2005), none has provided a quantitative measure on slant in education content. Previous studies on history textbooks (Ye, 2016[2]; Sneider, 2013; Cantoni et al, 2016) rely on human interpretation which inevitably involves a certain degree of subjective judgment. The quantitative approach used in this paper can mitigate the potential human bias in interpretation. By focusing on text subjectivity, we also bypass the need to find a language benchmark that is required for measuring bias. For example, current state-of-the-art method used to measure media bias in the U.S. requires the researcher to know how Democrats and Republicans talk in order for the algorithm to "learn" the partisanship rhetoric. These reference is often unavailable in authoritarian regimes or for text that do not have readily comparable references such as history textbooks or statement by the judges. Although this paper focuses on the cross-region difference in subjectivity, the approach is also fruitful in analyzing changes of subjectivity over time.

We use a number of metrics to measure degree of subjectivity. First, we count how often adjectives and four-character idioms (Chengyu) are used relative to the character length. Four-character idioms are unique in East Asian languages and often carry strong philosophical connotation and convey a sense of lecturing. The rationale of this measure comes from the observation that provocative and inflammatory language are often aided by use of adjectives. Adjectives are usually difficult to verify and quantify (what does a "brutal" war mean?). This means that it is difficult to dispute with the information sender. The sentiment analysis literature has also found that the presence of adjectives is strongly predictive on whether an article belongs to an opinion piece (Hatzivassiloglou and Wiebe, Wiebe, Bruce, & O'Hara 1999; Bruce & Wiebe 2000).

Second, we count the number of positive and negative words and calculate the ratio of positive to negative words. The use of positive and negative words directly expresses the sender's opinion on a certain historical episode or person. Several studies in Psychology have also found that pleasant and unpleasant adjectives are better memorized than neutral adjectives(Herbert et al, 2007; Kissler et al, 2009), suggesting that the use of emotional words could capture readers' attention. The ruling regime has an incentive to emphasize the policy success and underplay the mistakes so the positive-to-negative ratio on specific historical episodes and time period is

[2] The study examines if there are opposing historical accounts in Chinese and Taiwanese textbooks.

expected to vary on certain politically-sensitive episodes.

Third, we count the number of times entities of important historical and bipartisan political significance are mentioned. Mao Zedong - the central figure of the Chinese Communist party from its founding to his death - is critical to the CCP's authority and presumably mentioned more in the mainland textbooks. Similarly, Chiang kai-shek - the Director-General of the Kuomintang (KMT) until his death in 1975 - is expected to be mentioned more in the Taiwanese textbook. To facilitate comparison across the textbooks, we calculate the ratio of CCP to KMT count.

The first three methods simply count number of times a specific phrase appears. To capture the ways important political entities are being described in the text in more granularity, we use word embedding to measure distance between these political entities and other adjectives. Word embeddings use numeric vectors to represent words, and this representation allows us to measure the distance between important political entities such as 'Chinese Communist Party' and other adjectives in different textbooks. The method originally introduced by Bengio et al in 2003, and the main idea is that words that tend to appear in similar contexts are likely to be related. To learn these word embedding, we use Word2vec (Mikolov et al (2012)), a state-of-art method in the field of natural language processing that is implemented in Python.

We find that adjective ratio, positive-to-negative word ratio and the ratio of CCP to KMT entities mentions are both higher in the mainland textbooks than that in the other two regions as well as SMC. Mainland textbooks also have a much higher ratio of positive-to-negative words in post-1949 history (1949 being the year the People's Republic of China was established). In addition, the CCP political entities tend to have stronger association with positive words in mainland textbooks. Overall, the result confirms the hypothesis that the history textbooks in mainland China exhibit stronger degree of subjectivity across the three regions using our measures. We also find the adjective ratio is higher in the modern textbooks than the old textbooks in the mainland, and that the positive-to-negative words in post-1949 history registered a considerable increase over time also. In contrary, textbooks in Taiwan have become less subjective over time by our measure.

Psychologists and marketeers have long noted that messages, despite devoid of useful information, can improve the audience's assessment of the persuader's issue or product by triggering associative thinking or associative feeling via connotation of the message (Tversky and Kahneman 1982). Such association operates through a cognitive process that can take advantage of existing analogy an audience may already have in mind or by creating the relevant analogy by advertising attributes associated with that desired analogy (Mullainathan et al 2008). Our finding suggests that authoritarian regimes use messages with emotional connotation to trigger desired association with the regime. As another but extreme example of this tactic, propaganda

from North Korea often involves use of superlatives and hyperbole. The Korean Central News Agency - the state-controlled office in charge of communicating with the outside world- regularly promised "strikes without mercy by the steadfast and ever-vigilant People's Army.", and referred the South Korean government as "wicked running dogs" and "loud-mouthed gangsters."

This paper contributes to the growing literature across various social science disciplines in using data-intensive text analysis techniques to elicit political content (for example: Gentzkow and Shapiro 2010, Tumasjan et al 2010, Beauchamp, 2016; Roberts et al 2016). By pointing out the role of subjective sentiment in persuasion, it also expands the literature on strategies of persuasion. Much of the focus of this literature is on informative communication and understanding the equilibrium outcome (Kamenica and Gentzkow, 2011), with a few exception that emphasizes on behavioral limitation of analyzing the given information(Mullainathan et al, 2010). Recent work in political economy has also analyzed authoritarian regimes' incentive to homogenize the population by shaping the views of their citizens (Alesina and Reich, 2013; Cantoni et al, 2017), to which we provide indirect evidence which is manifested in educational material.

2 History Textbooks in Mainland China, Hong Kong and Taiwan

Our contemporaneous textbook sample consists of 3 history textbooks from the mainland, 2 from Hong Kong and 3 from Taiwan. They are purchased through an online shopping platform *Buyippee*. The textbooks are all used in local high schools as of 2016. Students studying these textbooks are generally between the age of 15 and 17. We sought to select textbooks that are most widely used in local high schools. To digitalize the textbooks into text files for computation, a undergraduate research assistant is hired and a combination of voice typing and manual entry methods are used. For all textbooks, we only use the history after the Opium War (A.D. 1842). In addition, we only use the main text portion and ignore all text in captions, supplementary information and appendix.

History textbooks in mainland China are regulated by the central government. The three approved versions are: Renmin, Renjiao and Xueli. Provinces can choose one of the three approved versions for their high schools [3]. The organization of the content is very similar across the different versions: there are three books in each version and each book has different emphasis. The first book focuses on the historical events and present them in a generally chronological order. The second book focuses on the economic development in the corresponding time period, whereas the third book focuses on cultural and technological progresses. In certain sections of the books, information that is not directly related to China is also presented.

[3] http://godfreyxu.github.io/2013/01/30/high-school-history-textbook-version-of-provinces-and-cities.html

For example, the different political systems used in other parts of the world as well as western literature and science advancement at the concurrent time period are discussed. We exclude these discussion in our text analysis to facilitate valid comparison across regions. In short, we only retain portions of the text that are directly relevant to mainland China. In addition, the curriculum is separated into mandatory and elective portion, and we use the mandatory potion only.

In both Hong Kong and Taiwan, the Education Bureau issues guidance for private publishers, and the schools have the discretion to decide which publisher's textbook version to use. In Taiwan, there are seven popular versions, and our sample consists of three of the most widely used: Kangzi, Nane, and Lungtun. Kangzi is most popular textbook and used by 45% of the schools, followed by Nane (17%) and LungTun (16%) [4]. The history curriculum is organized into four sections with a respective focus on Taiwan History, Chinese History, Ancient World History and Modern World History[5] and the events are presented in chronological order. We use the text about Chinese History in the second section only.

Our Hong Kong textbooks sample includes two versions from the following publishers: Manhattan Press, and the Modern Educational Research Society Limited. Both versions are among the most commonly used in Hong Kong [6]. Manhatten Press separates the curriculum into six books while Modern Education has 4 books, but the content and emphasis is very comparable between the two with the content of each section overlaps significantly with each other. The content is presented in a generally chronological order.

To measure over-time change in subjectivity, we also acquired two old textbooks prior to major reform in both mainland China and Taiwan. The old Taiwanese history textbook was use between 1983 and 1999 before the textbook market opened up for private publishers. It was standardized by the Ministry of Education so all high schools in Taiwan adopted this version. We do not have the mainland history textbook for the corresponding years, but we have a version used in 2003, just before a major reform took place in the mainland in 2004. The old textbooks allow us to measure the change in degree of subjectivity prior and after reform in the respective regions.

In addition to the secondary school textbooks used in the three regions, we also obtained a translated copy of the book "The Search for Modern China" (SMC) written by the British-American Scholar Jonathan Spence[7]. SMC has been assigned as a reference in many Chinese

[4]The figure comes from a study that sampled 313 schools in 2009 (Mao, 2013)

[5]https://www.sanmin.com.tw/learning/public/data/course

[6]There is no formal study that examines the fraction of schools using which textbook versions, but Manhattan Press, Modern Educational, Hong Kong Educational and Ling Kee are considered to be the four most widely used versions: http://www.com.cuhk.edu.hk/ubeat_past/051170/64.htm

[7]The book is translated by a Taiwanese.

history classes taught in the universities in the United States. Since this book is relatively free of a political agenda, it serves as a good comparison for the textbooks in the three regions despite the many difference in scope and depth. Specifically, on top of the general history, SMC also covers many areas of Chinese history from 1685 to 1989 which are traditionally outside the scope of high school textbooks such as the lives of peasants or people who are lesser known in history.

3 Descriptive Statistics

As part of preprocessing, we converted all text into simplified Chinese. We also removed all numbers, punctuation and non-Chinese characters. Unlike English, there is no white space in between words in Chinese so we need to segment the text into meaningful words and phrases. To do that, we use Jieba, a Python library, that is designed to segment Chinese sentences into individual phrases by finding the most probable combination based on the word frequency[8]. After segmenting the sentences, we removed the stop words from the text[9].

Table 1 provides the descriptive statistics in each version. The first column indicates the raw total character count in each version (by combining individual books for each version). The total character count varies considerably across regions though not as much within. Textbooks in Hong Kong are on average the longest, followed by that of mainland and Taiwan. SMC is considerably longer than our textbook sample. Column (2) - (4) indicate the number of bigram, trigram and quadgrams that Jieba detected.

In all versions, the number of trigrams and quadgrams identified in the dictionary is considerably less than that of the bigrams. Importantly, in all Mainland versions, the number of gradgrams exceeds the number of trigrams while the reverse is true for Hong Kong and Taiwan textbooks. Quadgrams are usually colorful adjectives, and many of which are 4-character idioms, which sometimes contain a fable with a practical lesson. This provides the first indication that mainland textbooks contain more subjective elements [10]. Column (5) and (6) shows the character count on pre- and post- 1949 content. Character count in Post-1949 history is less than that of pre-1949 in all textbooks. On average, the fraction of post-40 to pre-49 character count is lowest in the Taiwan textbooks and highest for mainland textbooks.

[8]We supplemented a list of names to the jieba dictionary to help jieba distinguish names from other parts of the sentence.

[9]We use the list of Chinese stop words from an online resource: `https://gist.github.com/dreampuf/5548203`

[10]In the sentiment analysis literature, Pang et al. (2002) also find evidence that higher-order n-grams are useful features in predicting opinionated piece. They report that unigrams outperform bigrams when determining determine whether a movie review is positive or negative

3.1 Major Historical Episodes

We extract from the text several major historical episodes to examine whether the treatment and emphasizes varies across textbooks on individual events[11]. We further group them into pre- and post-1949 events. Event such as the Cultural Revolution is expected to receive larger differential emphasis across the region's textbooks whereas event such as Foreign invasion during the Qing Dynasty should be relatively uniform.

The top panel in table 2 shows the number of words and the ratio to the character length in the four major historical episodes prior to 1949: Foreign Invasion (1842 -), XinHai Revolution (1911), Second Sino-Japanese War (1937-45), Chinese Civil War (1945 - 1949). Foreign Invasion refers to all of the wars involving a foreign power since the Opium War but excluding the second Sino-Japanese War. Xinhai Revolution includes description on Sun yut-chan, and all immediate events leading up to the Xinhai Revolution. The second Sino-Japanese War refers to the military conflict between China and Japan from 1937 to 1945. Finally, the Chinese Civil war includes all events surround the conflict between the Chinese Communist Party and Kuomintang after the surrender of Japan in 1945, but not any previous conflicts. We avoid capturing stand-alone sentences that may be relevant to the event, and capture a block of text whenever possible. We do not include text from Search of Modern China in this analysis because its writing styles makes it difficult to separate out the specific portion that is exclusively relevant to the major historical episodes.

Foreign Invasion has the largest word count among the pre-1949 events for 5 of the 8 versions in all regions. In Mainland Renjiao's version, Civil war has the largest word count, whereas in the Taiwan Kangxi and Lungtun version Xinhai Revolution has the largest word count. The word count varies significantly across versions within the same region. For example, Mainland's Xueli only has 618 characters on the Sino-Japan war whereas Renmin has 3293 characters. In terms of length ratio, the three Taiwan versions have the highest ratio for the Civil War. On Foreign Invasion, the two Hong Kong versions have the highest ratio across all 3 regions

The bottom panel of table 2 indicates the character count and the text ratio (text count/length of book) on three post-1949 historical episodes: the Great Leap Forward (1958 - 61), Cultural Revolution (1966-76), the Reform and Open Policy (1976 - current, encompassing the Tiananmen Square Protests in 1989). The reform and open policy receives the most character count across all versions even though the text ratio varies. On the two politically sensitive events, the Great Leap Forward has the lowest ratio in mainland's Renjiao, whereas the Cultural Revolution has the lowest ratio in mainland Xueli. They the highest text ratio in the two Hong Kong versions. Interestingly, on the coverage of Reform and Open policy, Taiwan Lungtun has the highest ratio

[11]Since mainland versions have different emphasis in each book, we exclude text in the books on economics and culture and technology and only use the text in the general history.

and the lowest goes to Taiwan's Nane, suggesting considerable variation in event emphasis within Taiwanese versions. The Tiananmen Square Protests is noticeably absent in all mainland textbooks but are mentioned in all versions in both Hong Kong and Taiwan.

4 Subjectivity Analysis

We construct four metrics to measure subjectivity. First, we count the number of times an adjective is used and calculate the ratio of adjectives to the total word count. Work on subjectivity detection revealed a high correlation between the presence of adjectives and sentence subjectivity (Hatzivassiloglou and Wiebe, 2000). Knowing the affective aspect of the opinions like "happiness" or "mood" is also important in understanding government's objective so for our second measure, we calculate the positive to negative word ratio using a sentiment dictionary. Third, in order to assess how often a figure of partisan significance is mentioned, we count the number of times important political entities such as Mao Zedong and Chiang Kai-shek is mentioned. Forth, we use word embedding to calculate distance between relevant politically-relevant entities and other adjectives.

4.1 Adjective Ratio

To determine whether a phrase is an adjective, we wrote a script in Python to automatically look up all the phrases in the Chinese dictionary pre-installed on a Mac notebook[12] . A phrase is considered as an adjective if the word definition contains the character 形 (English: adj.) and a four-character idiom if the character 成 is present. In addition to adjectives, we also examined the presence of adverbs. A word/phrase is considered as adverb if the character 副 is present. Table 3 lists the top 5 bigram and quadgrtam adjectives and their respective frequencies in each textbook version. 独立自主 (Act independently and of one's own initiative) is the most common quadgram for all 3 mainland versions. All popular quadgrams in mainland versions have a positive connotation. For the two Hong Kong versions, 内忧外患 (Domestic trouble and foreign invasion) is among the top 5 quadgrams. 革命 (Revolution/Revolutionary) and 经济 (Economy/Economic) are the two most popular bigrams in versions of all regions. Perhaps surprisingly, 民主 (Democracy/Democratic) is the third most commonly used bigram adjective

[12] `jieba.posseg.cut` in the Jieba module returns part of speech tagging but the result is not satisfying. For example, the noun 蛮夷 means savage or barbarian people. Using `jieba.posseg.cut` will not detect the negative connotation of this phrase because it will only return "noun" for the phrase. Ideally, we want the segmentation algorithm to tell us that 蛮夷 is a subjective noun. While this is not currently available in jieba, using the "full" method allows us to separately look at each possible word combination to determine if any of the phrase has an adjective component. An example of jieba's full segmentation method follows. The sentence 我来到北京清华大学 can be segmented into 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学 . The phrase 清华大学 is inevitably counted twice because of the different possible combination within the phrase. However, this allows us to separately examine the individual word within 清华大学 .

in all mainland versions[13].

Figure 4 and 5 plot the adjective and adverb ratio for each textbook version. The three mainland versions have the highest adjective ratio among all textbooks. On average, the adjective ratio in mainland textbooks is more than 1% higher than Hong Kong and Taiwan textbooks. As a reference, SMC has a similar adjective ratio as the Hong Kong and Taiwan textbooks. The adverb ratio is slightly higher for Taiwan textbooks relative to Hong Kong and mainland versions, but is similar with that of SMC. The ratio is similar between Hong Kong and mainland versions. The difference in adverb ratio is smaller than that of the adjective ratio, with the difference between the largest and smallest about 0.7%. This suggests that the use of subjective element is more subtle in mainland textbooks than popular writings. To the extent that adverbs are more common in day-to-day speech than formal writing, this could be due to the difference in writing style that demands more formality in textbook texts.

Figure 6 and figure 7 plot the adjective ratio for the 4 pre-1949, and 3 post-1949 historical episodes in each version. In the civil war portion, the adjective ratio of the three mainland versions stands out compare to versions in other regions. Across all versions, the adjective ratio is highest for Xinhai Revolution compared with the other pre-1949 episodes. Among post-1949 historical episodes, mainland versions have a higher adjective ratio in the Reform and Open portion. Adjective ratio is also higher in the Great Leap Forward portion in mainland versions though the variation is also higher among different versions and across regions[14].

Figure 8 plots the adjective ratio of the pre- and post-1949 content. The pre-1949 adjective ratio of the mainland textbooks are on average highest among all versions, even though two of the mainland versions appear to be not far from SMC and the Hong Kong textbooks. More notedly, the adjective ratio of all versions increases or remains flat from pre- to post-194, but the average increase in mainland textbooks is much larger when compared with the other books.

4.2 Text Sentiment

Text sentiment - whether it is positive, negative, or neutral - informs the audience on the author's opinion. Studies in political communication and marketing have found that text with strong valence tend to be more memorable even though the effect of positive and negative messages tend to be asymmetric[15]. To determine the sentiment implied in the text, we assign a phase

[13]The bigram adjectives are commonly regarded as nouns as well. This can make our result difficult to interpret. We verify that the overall pattern of our results remain unchanged after manually removing these noun/adjectives in the ratio calculations.

[14]Due to the low total character count in Great Leap in mainland versions, the ratio is not the best measure of subjectivity.

[15]Lau et al. 2006 find that negative political advertising can stimulate knowledge about the campaign. In Lee et al., 2007 and Chevalier and Mayzlin 2006, the researchers found that negative product reviews can change initial attitudes and influence purchase decision

as either positive or negative using a binary sentiment Chinese dictionary available online[16]. The dictionary consists of 4570 positive and 4374 negative phrases, and are not restricted to adjectives. Some of these phrases are subjective nouns such as 爱戴 (love and support) and 褒扬 (praise). While the set of reference phrases is limited, it provides a reasonable starting point. Phrases not found in the dictionaries are ignored in our calculation. We then performed the same preprocessing on the text including the removal of stop words and punctuations.

Table 8 and 9 show the top 5 most used positive and negative phrases. For both positive and negative phrases, the number of phrases that are present in the dictionary is higher in mainland textbooks. 严重 (serious) is the most common negative phrase, and it is among the top 5 in all versions. Present in 8 versions, 新 (new) is the most common positive phrase. There is significant overlap among the most frequently used positive and negative phrases in textbooks of the same region. For example, the top 5 positive phrases in the mainland Renjiao and Xueli versions are identical, and both 专制 (dictatorial) and 严重 (serious) are present in all three mainland versions. 3 positive phrases, and 4 negative phrases shared the top 5 between the 2 Hong Kong versions. For the case of Taiwan, Kangxi and Lungtun share more commonly used phrases, but less so with Nanye.

Table 10 reports the number of positive, negative phrase, and the positive-to-negative ratio for each version. Mainland textbooks have the highest higher positive to negative words ratio, followed by Hong Kong and Taiwan. In particular, Renmin's ratio is about 2 times higher than the ratio of SMC, which has the lowest ratio of all. On average, the positive-to-negative ratio is slightly higher in Hong Kong than in Taiwan.

Figure 10 and 11 plot the results of the positive to negative word ratio in the major historical episodes. The ratio is similar across all versions for Foreign Invasion. There is more variation between versions for the other three events. The variation within mainland versions is particularly large for Xinhai Revolution: Renjiao has the lowest ratio whereas Renmin has the highest. Within Hong Kong versions, the ratio vary much less across all 4 episodes compare to the other two regions. Shifting to post-1949 episodes, the ratio is considerably higher in all three mainland versions for the Reform and Open section: it is about 2 times higher than Hong Kong and about 3 times than Taiwan. The ratio is similar across versions and regions in Cultural Revolution and Great Leap Forward.

Figure 9 plots the pre- and post-1949 positive to negative word ratio for each version. In the pre-1949 period, the ratio is similar across all textbooks at around 2 to 3 positive to 1 negative word. However, mainland versions have the largest increase from pre- to post- 1949 with about 5 to 7 positive to 1 negative word in the post-1949 period. The average ratio for Hong

[16]The dictionaries are available at: `https://github.com/Fansion/CCUED/blob/master/README.md`. The dictionary has the drawback of not having many idioms.

Kong versions has a slight increase, while Taiwan has a small decrease on average. SMC has a roughly similar ratio pre- and post-1949. The result strongly supports the notion that the mainland textbooks portray a more positive image of the regime compared with the textbooks of the other two regions.

4.3 Political Entities

The number of times a political entity is mentioned is a direct indication of the political entity's importance in history, but it can also suggest the political entity's significance specifically to the regime. For example, "Mao Zedong", as well as "Chinese Communist Party" (CCP) are expected to be more frequently cited in mainland textbooks because of their significance in mainland politics. In contrary, "Kuomintang" (KMT) is expected to mention more in Taiwan textbooks.

We count the number of times several important CCP and KMT entities are mentioned in each version, and then calculate the ratio of CCP to KMT entities mention by considering the following list of historically significant individuals and entities: Mao Zedong, the Chinese Communist Party, Karl Marx, Chiang Kai-Shek[17] , Sun Yat-Sen, the Kuomintang. The first three are considered to be CCP and the last three KMT entities. The ratio is presented in figure 3. Mainland versions have the three highest CCP to KMT ratio among all versions. In particular, the Renmin version mention a CCP entity in about 3.5 to 1 ratio, while SMC has a 2 to 1 ratio. Hong Kong and Taiwan textbooks have a ratio between 1 to 2 with Taiwan version on average the lowest among the three regions. It is worthwhile to note that the CCP to KMT ratios can be interpreted as either a subjectivity or bias measure based on our definition. Opinion representing the regime or political entity can be treated as subjective content. In contrary, factual information in support of the political entity can be considered as bias. [18]. We made no distinction here.

Table 7 presents the number of mentions for each of the 6 individuals and entities. The contemporary mainland versions mention Chiang less than 15 times but Mao is mentioned more than 45 times. The opposite is true in Taiwan versions. Marx is mentioned less than 5 times in all Taiwan versions but at least 19 times in the mainland versions. Hong Kong versions mention Marx 9 times on average. The number of mentions of CCP is comparable between mainland and Taiwan versions, with the average being 113 times in mainland versions and 103 times in Taiwan versions. Sun also appears to receive relatively equal emphasis between mainland and Taiwan versions. Both region mention Sun about 30 times on average. In comparison, Hong Kong versions mention Sun 65 times. Interestingly, KMT is mentioned more times in mainland versions that in Taiwan versions. On average, mainland versions mention KMT 96 times while

[17]The Taiwanese versions address Chiang Kai-Shek using his adopted name "Zhongzheng", and in many cases in the old Taiwanese textbook, the honorific "Chairman Chiang" is used.

[18]By counting the number of times a think tank is cited in the newspaper, Groseclose and Jeffrey Milyo (2005) have considered a similar idea to capture liberal bias in the media.

Taiwan versions 47 times. Taken together, the count suggest that Marx, Mao and Chiang Kai-Shek receive more drastic difference in treatment between mainland and the Taiwan, whereas CCP, KMT and Sun are relatively comparable.

4.4 Word Embedding

Cutting the text by time period and historical episode is a crude way to assess how adjectives are associated with the regime. To understand how a political entity is being portrayed in more granularity, we can study the word embedding of the political entity. In Linguistics, word embeddings aim at quantifying semantic similarities between words based on their distributional properties. The basic idea is that words with similar distributions of other words have similar meanings. Mathematically, a word embedding $W : words \rightarrow \mathbb{R}^N$ is a parameterized function mapping words in some language to high-dimensional vectors. Using this framework allows us to capture the sentiment surrounding the political entities of interest by looking at the distance with their surrounding adjectives in the high-dimensional embedding space trained by the history textbooks from each region.

We use Word2vec (Mikolov et al, 2013), which is a group of related models that are used to produce word embeddings, to consider the associations between entities of interest (such as 'Taiwan', 'Chinese Communist Party') and words with emotional connotation. Word2vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text. One type of model within Word2vec is called Continuous Bag of Words (CBOW) (Mikolov et al, 2013). It can be used to predict co-occurrence relationships using the conditional probability of observing the target word given the input context words. Context words is represented by multiple words for a given target word. For example, Word2vec treats "The", "cat", "over", "the", "puddle" as context and and predicts the target word "jumped". The training goal of the CBOW model is to arrive at vector representations of words that best predict the target word. Formally the objective function is given by:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} log\ p(w_t|w_C) \tag{1}$$

where θ represents all the variables we optimize. w_t denotes the target word, and w_C denotes the context words. t denotes the training step. Denote V as the vocabulary in the text. The conditional probability of the target word can be represented by a softmax function, which uses a neural network structure to learn the parameters:

$$p(w_t|w_C) = \frac{exp(u_t^T v_{w_c})}{\sum_{w_i \in V} exp(u_w^T v_{w_c})} \tag{2}$$

u_w and and $v_w$are two representations of the word w. u_w comes from rows of the input to hidden weight matrix, and v_w comes from columns of hidden to output matrix. The inner product $u_t^T v_{w_c}$ computes the log-probability of word w_c, which we normalize by the sum of the

log-probabilities of all words. The goal of the algorithm is to learn the weights in the input to hidden layer, and hidden layer to the output matrix[19]. (A more detailed explanation of the CBOW model and neural network learning can be found in Rong (2014)).

To measure the distance between words, we calculate the cosine distance between the word vectors in the high-dimensional embedding space. The cosine distance between vector A and B is defined as:

$$\frac{A \cdot B}{||A||\,||B||} \tag{3}$$

The resulting similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality. We can then examine the distance between important political entities to other phrases.

Rather than treating each version separately, we combine individual textbook versions into a corpus for each region and train our model based on region-specific text. This is because we want to maximize statistical power in explaining cross-region difference in language use. We trained each model with an embedding dimension of 500 and with context size of 6 using a python library `gensim`, which has built in a CBOW model. We searched for the closest adjectives around the important political entities in each embedding space. The result for each region is presented in tables 4, 5 and 6. Positive adjectives such as 拨乱反正 (bring order out of chaos) and 解放思想 (liberate thoughts) are found to be closest with Mao Zedong in the Mainland textbooks, whereas more neutral phrases such as 整肃 (enforce) are among the closest in Hong Kong and Taiwan textbooks. On Chiang Kai-Shek, positive phrases such as 取得胜利 (to get victory) and 固守 (defend tenaciously) are among the closest adjectives in Taiwan textbooks whereas negative phrase such as 全军覆没 (the whole army is wiped out) appears as close adjectives in Mainland textbooks.

We calculated the ratio of positive to negative phrase ratio using the top 20 surrounding adjectives on the 6 political entities. Table 11 shows that Mainland versions have higher positive-to-negative phrase ratio on Mao, Sun and Marx. On the other hand, Taiwan versions have higher positive-to-negative ratio only on Chiang. On average, Hong Kong versions have mild positive-to-negative ratio on all 6 entities.

We can visualize each embedding space trained by the model by using the t-Distributed Stochastic Neighbor Embedding method (t-SNE) (Maaten et al, 2008). This is a dimensions reduction technique that is particularly well suited for the visualization of high-dimensional datasets such as ours. The idea is to embed high-dimensional points in low dimensions in a way that respects similarities between points[20]. By visualizing them in a 2-dimensional plane, we can

[19]This can be achieved by the gradient descent method using a random initialization.

[20]t-SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. Suppose there are only two high-dimensional objects, x_1 and x_2.The conditional

illustrate the semantic distance between Mao, CCP, Marx, Chiang, KMT, and Sun in each of the mainland, Hong Kong, and Taiwan textbooks. Figure 17 shows the embedding space of mainland Chinese textbooks. While the relationship among Marx and CCP and Mao is like that of Sun and Chiang and KMT, the two clusters are far from each other suggesting that . Figure 16 is the visualization of the embedding space of the Hong Kong textbooks. It shows three clusters, CCP and Marx as one, Mao as another itself and KMT, Chiang and Sun as the third. Fig 17 shows that Mao, CPP, KMT, Chiang and Sun are clustered and Marx alone is far away.

5 Prior and After Reform

History textbooks in both mainland China and Taiwan have undergone significant revision in the last two decades. Before 1996, Taiwan's history textbooks were designed and published by the National Institution for Compilation and Translation under the Ministry of Education (MOE). Schools were required to use the nationally standardized textbooks. Since 1996, the MOE has opened up the textbook market for private publishing companies, which means that there is more variation in content across textbook versions and deviation from the official interpretation of history is permitted. Schools can choose the textbooks that they determine as apt to the curriculum. In contrary in mainland China, a nationwide education reform was proposed by the central government in 2001 with an explicit goal to instill patriotism and love for social socialism in students. [21]. The reforms in these two regions have adopted opposite direction and should have diverged implication on text subjectivity. Specifically, text subjectivity is expected to increase in mainland textbooks and decrease for the Taiwan textbooks after the reform. We test these implications next.

We first examine the summary statistics. The bottom 2 rows of table 1 print the book length and the portion allocated to pre-1949 and post-1949 history for the old versions. The old versions of both the mainland and Taiwan textbooks have larger word count than the respective modern versions. The post-1949 portion of the old Taiwanese version is significantly smaller compared to the modern version with only one paragraph allocated to post-1949 mainland China. Sim-

probability $p_{2|1}$. The goal is to learn a 2-dimensional projection y_2 and y_1 that reflects the similarities between x_1 and x_2 as well as possible. As with their high-dimensional counterparts, the similarity of the projection points can be represented by conditional probability. The observation is that if the map points y_1 and y_2 correctly model the similarity between the high-dimensional data points x_1 and x_2, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. The algorithm that minimizes the sum of Kullback-Leibler divergences over all data points using a gradient descent method with respect to y.

[21] From Appendix C of Cantoni et al(2016), in the Ministry of Education's "Framework for Basic Education Reform" (2001), a new curriculum should meet the following objectives (in the order of appearance in the original document): it should reflect the times, and make students patriotic, communitarian, [and] love socialism. Students should inherit and carry forward the great traditions of the Chinese nation and its revolution; and be equipped with an awareness of the legal system under a socialist democracy. The new curriculum should promote compliance with national laws and with societal ethics, and gradually form in students a correct world view, a correct view of life, and a correct value system.

ilarly, the old mainland version allocates a larger fraction of its total word count to pre-1949 history compared to the modern versions: the ratio of word count of pre- to post-1949 is over 2 in the old version and less than 1.5 in the modern versions on average.

As shown in figure 4, the aggregate adjective ratio increases from 0.52 to 0.55 for the mainland textbook. The adjective ratio of the old Taiwanese version is about the same with the average of the modern versions. Figure 12 plots the adjective ratio of the old version with the average of the modern versions for each historical episode. Since the text size of individual events is rather small and changes considerably before and after event, we warrant caution in drawing conclusions on the change in subjectivity for individual events. Nevertheless, the adjective ratio appears to be relatively similar in the four pre-1949 episodes. Among the post-1949 episodes, there is a noticeable increase for the Great Leap Forward but decreases in Cultural Revolution in mainland textbooks. For the Taiwan versions, the adjective ratio appear to stay relatively similar or decrease slightly for the four pre-1949 episodes.

Figure 13 (**??**) plots the adjective ratio (positive to negative phrase ratio) for the pre- and post-1949 history of the old versions and the average of the modern versions by region. The increase in adjective ratio from pre- to post-1949 in the old Taiwanese version is much larger than that in the modern Taiwan version. A similar increase in adjective ratio from pre- to post-1949 can be observed in the old mainland textbook. The increase is parallel to the increase of the modern mainland versions[22]. In terms of the positive to negative phrase ratio, the average of modern mainland versions has a much higher ratio than the old mainland version, suggesting that there is more emphasis on the positive aspects since 1949[23].

From table 7, CCP is mentioned 152 times in the old mainland version and 74 times in the old Taiwanese version. Interestingly, KMT is mentioned more in mainland than in Taiwan: 116 times in the old mainland version and 32 times in the old Taiwanese version. Chiang is mentioned about equal number of times with 41 and 42 between the two, as well as Sun, mentioned 37 and 39 times in the old mainland and the old Taiwanese versions. After the reform, Chiang is mentioned in a 2 to 1 ratio in modern Taiwan and mainland textbooks. In contrast, Mao is mentioned 62 times in the old mainland version and only 2 times in the old Taiwanese version. After the reform, Mao is mentioned in higher frequency but Marx has remained scarcely mentioned in modern Taiwan versions. This suggests that Mao and Marx are more polarizing than Chiang and Sun before the reform. After the reform, Chiang became more polarizing but Mao less so.

[22] Figure 13 calculates the adjective ratio using only first book of the mainland textbooks. As the second and third book focuses on economic and cultural aspect of the corresponding time period, it is less comparable with the old textbook **??** shows the adjective ratio using all 3 books of the modern mainland versions.

[23] For Taiwan textbooks, since the text size on post-1949 in the old Taiwanese version is very small, the low positive to negative phrase ratio can be misleading.

6 Writing Style

Can the result be attributed to difference in writing styles across regions? That is, do mainland Chinese tend to use more adjectives, and emphasize the positives in their writing? While this is a possibility, and indeed the difference in speech and informal writing style is large across the regions especially between Hong Kong and the other two regions[24], formal writing in Chinese is relatively similar. Nonetheless, to evaluate this possibility, we select 3 representative newspapers of the three regions and quantitatively examine their writing styles including the adjective ratio and the positive to negative phrase ratio.

The three newspapers we selected have reasonably large circulation. We focus on the section on economy only since the economy section should be relatively free of subjective interpretation and allows us to focus on writing style. We avoid choosing a newspaper that operates under direct government influences such as the state-owned newspaper. The newspaper sample are Shanghai Morning Post (新闻晨报) circulated mainly in Shanghai, Oriental Daily (東方日報) in Hong Kong and Liberty Times (自由時報) in Taiwan. The news content is available from Wisers, a Hong Kong private company. We use all news reports in year 2016[25].

As shown in the top panel of table 12, the adjective ratio in the newspapers is higher than that in the history textbooks in all 3 regions, suggesting that the writing style of newspapers tend to be different from that of history textbooks which might be attributed to difference in how formal and rigid of the text requirement. More importantly, we found that Taiwan newspaper has the highest adjective ratio, followed by that of Mainland and Hong Kong. The positive to negative phrase ratio of the three newspapers is presented in the bottom panel of table 12. With a range of 3.5-4.3, the positive to negative phrase ratio of the newspaper is not too far from that of the history textbooks. Across the newspapers, Liberty Times has the highest positive to negative phrase ratio, followed closely by Shanghai Morning Post. Oriental Daily has the lowest with approximately 3.5 positive to 1 negative phrase. The result suggests that adjective ratio as well as the positive-to-negative phrase ratio do not tend to be higher in the mainland writings. This supports our hypothesis that the use of adjective and phrase with emotional connotation is a manifest of persuasion.

7 Conclusion

This paper provides the first quantitative analysis on history textbooks by focusing on text subjectivity. We have shown that mainland Chinese textbooks exhibit a stronger degree of subjectivity by the following measures: the adjective ratio, the ratio of positive to negative phrase, the ratio of CCP to KMT entities mentions, and distance between political entities and adjec-

[24] The primary spoken language in Hong Kong is Cantonese while the other two regions primarily speak Mandarin.
[25] We exclude all non-news articles including columns that tend to represent one's opinion.

tives. The result broadly confirms the notion that the authoritarian government has a stronger incentive to present subjective content to influence the political views of the students. In addition, by studying the textbook before and after reform in mainland China and Taiwan, we show that looser government control of education content led to a more objective evaluation of history.

By examining subjectivity rather than bias, this paper focuses on an aspect of persuasive communication that has received scant attention in the social science literature. The methods demonstrated in this paper can also be used to analyze the subjectivity of other politically-relevant content such as speeches by politicians and statements issued by judges. There are important caveats, however. First, our approach is comprised of several individual measure, so it could be difficult to access the overall subjectivity when different measures give contradicting results. Second, our method only considers adjective, and does not capture other subjective elements such as subjective nouns. Previous study in Natural Language Processing has found that certain extraction pattern can be leveraged to identify subjective words (for example, the pattern "expressed *object*" often extracts subjective nouns, such as "concern", "hope", and "support"), and some bootstrapping algorithms can automatically generate these extraction patterns(Riloff et al., 2003). We have not considered this possible extension in this paper.

How attitudes and opinions can be changed by text subjectivity is unanswered in this paper. An important question remains: while the intent of the government should be explicit, students' political views appear to be be still influenced by the curriculum (Cantoni et al 2016). Does this channel pf persuasion operate through cognitive or rational processing, or a combination of both? While this paper has illustrated the possibility of cognitive processing, future research can explicitly examine how text subjectivity as well as information on senders' intention influences opinion.

References

[1] A. Alesina and B. Reich. Nation building. 2013.

[2] M. W. Apple and L. K. Christian-Smith. The politics of the textbook. 1991.

[3] R. Batra and D. M. Stayman. The role of mood in advertising effectiveness. *Journal of consumer research*, 17(2):203–214, 1990.

[4] N. Beauchamp. Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[6] H. Bless, G. Bohner, N. Schwarz, and F. Strack. Mood and persuasion a cognitive response analysis. *Personality and Social Psychology Bulletin*, 16(2):331–345, 1990.

[7] H. Bless, D. M. Mackie, and N. Schwarz. Mood effects on attitude judgments: independent effects of mood before and after message elaboration. *Journal of personality and social psychology*, 63(4):585, 1992.

[8] L. Bursztyn and D. Cantoni. A tear in the iron curtain: The impact of western television on consumption behavior. 2012.

[9] B. J. Calder and B. Sternthal. Television commercial wearout: An information processing view. *Journal of Marketing Research*, pages 173–186, 1980.

[10] S. DellaVigna, R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya. Cross-border media and nationalism: Evidence from serbian radio in croatia. *American Economic Journal: Applied Economics*, 6(3):103–132, 2014.

[11] S. DellaVigna and M. Gentzkow. Persuasion: empirical evidence. *Annual Review of Economics*, 2:643–669, 2010.

[12] S. DellaVigna and E. Kaplan. The fox news effect: Media bias and voting. *Quarterly Journal of Economics*, 122(3), 2006.

[13] M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[14] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, pages 1191–1237, 2005.

[15] N. Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4):1359–1395, 2009.

[16] J. Lee, D.-H. Park, and I. Han. The effect of negative online consumer reviews on product attitude: An information processing view, 2008.

[17] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[18] D. M. Mackie and L. T. Worth. Processing deficits and the mediation of positive affect in persuasion. *Journal of personality and social psychology*, 57(1):27, 1989.

[19] C.-Y. Mao. 選什麼？如何選？為何而選？— 高中教師選擇歷史教科書之研究。教育科學研究. 教育科學研究, pages 123–144, 2013.

[20] J. D. Mayer, Y. N. Gaschke, D. L. Braverman, and T. W. Evans. Mood-congruent judgment is a general effect. *Journal of personality and social psychology*, 63(1):119, 1992.

[21] J. L. McCullough and T. M. Ostrom. Repetition of highly similar messages and attitude change. *Journal of Applied Psychology*, 59(3):395, 1974.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[24] H. F. Mosher and A. Banfield. Unspeakable sentences: Narration and representation in the language of fiction, 1984.

[25] S. Mullainathan, J. Schwartzstein, and A. Shleifer. Coarse thinking and persuasion. *The Quarterly Journal of Economics*, 123(2):577–619, 2008.

[26] T. O'Hara, J. Wiebe, and R. Bruce. Selecting decomposable models for word-sense disambiguation: Thegrling-sdm system. *Computers and the Humanities*, 34(1-2):159–164, 2000.

[27] M. T. Pham. Emotion and rationality: A critical review and interpretation of empirical evidence. *Review of general psychology*, 11(2):155, 2007.

[28] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. *Proceedings of the Seventh Conference on Natural Language Learning*, 2003.

[29] M. Roberts, B. Stewart, and E. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111:988–1003, 2016.

[30] X. Rong. word2vec parameter learning explained. *Manuscript*, 2014.

[31] F. R. Rosselli, J. J. Skelly, and D. M. Mackie. Processing rational and emotional messages: The cognitive and affective mediation of persuasion. *Journal of Experimental Social Psychology*, 31:163–190, 1995.

[32] A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758, 2003.

[33] D. C. Sneider. History textbooks and the war in asia.

[34] J. D. Spence. *The search for modern China*. WW Norton & Company, 1991.

[35] D. Strömberg. Mass media competition, political competition, and public policy. *The Review of Economic Studies*, 71(1):265–284, 2004.

[36] S. M. Tom, C. R. Fox, C. Trepel, and R. A. Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518, 2007.

[37] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[38] Z. Wang. National humiliation, history education, and the politics of historical memory: Patriotic education campaign in china. *International Studies Quarterly*, 52(4):783–806, 2008.

[39] J. M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.

[40] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.

[41] L. T. Worth and D. M. Mackie. Cognitive mediation of positive affect in persuasion. *Social cognition*, 5(1):76–94, 1987.

[42] D. Yanagizawa-Drott. Propaganda and conflict: Evidence from the rwandan genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994, 2014.

[43] A. Ye. Remembering different pasts: An analysis of history textbooks in mainland china and taiwan. 2016.

8 Figures and Tables

Figure 1: Screenshot of the dictionary application on mac that shows the word 'Important'

Figure 2: Screenshot of the dictionary application on mac that shows the word 'Very'

Table 1: Summary Statistics of Each Textbook Version

Region	Book	Book length	Bigram	Trigram	Quadgram	Pre-1949	Post-1949
Hong Kong	Manhattan	70627	26206	1317	1070	43851	28808
Hong Kong	Modern	91087	33866	1502	1356	59798	34451
Mainland	Renjiao	41544	16778	867	986	12144	7769
Mainland	Renmin	55484	22433	1226	1469	15003	13156
Mainland	Xueli	38908	15688	857	941	9959	8384
Taiwan	Nanyi	30792	11112	444	345	25027	8231
Taiwan	Kangxi	33640	12679	506	419	22346	11928
Taiwan	Lungtun	40457	15527	610	540	29357	12307
United States	Search of Modern China	410198	140591	7345	3962	235970	174094
mainland	old	56925	22746	1357	1416	41318	18101
taiwan	old	46940	16025	552	347	46668	274

The count represents phrases that are recognized in the Dictionary. For the mainland textbook versions, the pre- and post-1949 portion only includes the generic history portion, i.e. we did not include the portion with specific focus on economics and culture.

Table 2: Character Count and Ratio of Major Historical Episodes

	Hong Kong		Mainland			Taiwan			Mainland	Taiwan
	Manhattan	Modern	Renjiao	Renmin	Xueli	Nanyi	Kangxi	Lungtun	old	old
Foreign Invasion	12326	16865	2608	3526	2232	2414	2063	2095	7194	8369
	0.175	0.185	0.063	0.064	0.057	0.078	0.085	0.052	0.126	0.179
Xinhai Revolution	4448	4937	1496	1472	1184	1778	2846	2598	1764	1710
	0.063	0.054	0.036	0.027	0.030	0.058	0.061	0.064	0.031	0.037
Civil War	3460	4112	2822	890	1000	1907	2166	2631	3239	1367
	0.049	0.045	0.068	0.016	0.026	0.062	0.064	0.065	0.057	0.029
Sino-Japan War	3989	5991	1498	3293	618	1641	2016	1332	7879	4832
	0.056	0.066	0.036	0.059	0.016	0.053	0.060	0.033	0.138	0.104
Great Leap Forward	1484	1905	382	230	457	528	345	449	413	-
	0.021	0.021	0.009	0.004	0.012	0.017	0.010	0.011	0.007	-
Cultural Revolution	5834	4406	940	876	429	523	1249	625	2086	-
	0.083	0.048	0.023	0.016	0.011	0.017	0.037	0.015	0.037	-
Reform and Open	6303	7514	2308	2848	2774	1475	1644	4808	3295	-
	0.089	0.083	0.056	0.051	0.071	0.048	0.049	0.119	0.058	-

The second row indicates the ratio of the event length to the overall book length. Since the old Taiwan textbook contains a very short description of post-1949 history, we do not separate them into individual historical episodes.

Table 3: Most popular adjectives in each textbook

Renmin (人民)				Renjiao (人教)				Xueli(岳麓)			
Bigram	Count	Quadgram	Count	Bigram	Count	Quadgram	Count	Bigram	Count	Quadgram	Count
革命 (Revolution)	266	独立自主 (act independently and of one's own initiative)	11	革命 (Revolution)	158	独立自主 (act independently and of one's own initiative)	8	经济 (Economy)	139	独立自主 (act independently and of one's own initiative)	5
经济 (Economy)	156	实事求是 (be true to facts)	11	经济 (Economy)	140	百家争鸣 (contention and flourishing of numerous schools of thought)	7	革命 (Revolution)	128	百花齐放 (flourishing art and literature)	4
民主 (Democracy)	154	拨乱反正 (bring order out of chaos)	5	民主 (Democracy)	110	百花齐放 (flourishing art and literature)	7	民主 (Democracy)	124	深入人心 (strike a deep chord in the hearts of the people)	3
科学 (Science)	72	轰轰烈烈 (vigorous)	5	开放 (open up)	50	实事求是 (be true to facts)	4	科学 (Science)	47	当家做主 (rule the roost)	3
开放 (open up)	57	与时俱进 (keep pace with the times)	3	封建 (feudal)	46	欣欣向荣 (flourishing)	3	和平 (peace)	43	拨乱反正 (bring order out of chaos)	3
Nane (南一)				Kangxi (康熹)				Lungtun (龍騰)			
Bigram	Count	Quadgram	Count	Bigram	Count	Quadgram	Count	Bigram	Count	Quadgram	Count
革命 (Revolution)	89	有识之士 (man of insight)	2	革命 (Revolution)	111	有识之士 (man of insight)	3	革命 (Revolution)	95	贪污腐化 (corruption and degeneration)	5
经济 (Economy)	62	痛定思痛 (draw a lesson from a bitter experience)	2	经济 (Economy)	62	百家争鸣 (contention and flourishing of numerous schools of thought)	2	经济 (Economy)	65	国计民生 (national economy and people's livelihood)	2
国际 (International)	28	救亡图存 (save the nation from doom and strive for its survival)	2	国际 (International)	34	苛捐杂税 (exorbitant and multifarious taxes and levies)	2	主要 (primary)	42	救亡图存 (save the nation from doom and strive for its survival)	2
重要 (important)	28	改弦易辙 (change one's course)	2	开放 (open up)	32	内忧外患 (domestic trouble and foreign invasion)	2	国际 (International)	40	蒸蒸日上 (become more prosperous every day)	2
开放 (open up)	22	东山再起 (make a comeback)	2	重要 (important)	30	贪污腐化 (corruption and degeneration)	2	重要 (important)	38	独立自主 (act independently and of one's own initiative)	2
Manhattan (文津)				Modern Education (現代)							
Bigram	Count	Quadgram	Count	Bigram	Count	Quadgram	Count				
革命 (Revolution)	214	内忧外患 (domestic trouble and foreign invasion)	5	革命 (Revolution)	250	独立自主 (act independently and of one's own initiative)	14				
经济 (Economy)	178	实事求是 (be true to facts)	4	经济 (Economy)	224	内忧外患 (domestic trouble and foreign invasion)	6				
开放 (open up)	69	前所未有 (unprecedented)	4	开放 (open up)	98	有识之士 (man of insight)	6				
民主 (Democracy)	58	自负盈亏 (assume sole responsibility for profits or losses)	3	和平 (peace)	76	迫在眉睫 (extremely urgent)	4				
国际 (international)	53	国计民生 (national economy and people's livelihood)	3	国际 (International)	68	自负盈亏 (assume sole responsibility for profits or losses)	3				

Table 4: Mainland closest adjectives

Chiang (蒋介石)		Mao (毛泽东)		KMT (国民党)		CCP(共产党)		Sun(孙中山)		Marx(马克思)	
Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance
复新 (again)	0.7202	独秀	0.6510	正面 (positive)	0.6635	重新 (again)	0.5732	腐朽 (rotten)	0.5724	广泛 (extensive)	0.6825
(again)		(duk-sau) (name)		(positive)		(again)		(rotten)		(extensive)	
北新	0.7111	拨乱反正	0.5554	轰轰烈烈	0.5740	正面	0.5494	平均	0.4795	具体	0.6699
(North New)		(bring order out of chaos)		(vigorous)		(positive)		(average)		(concrete)	
全军覆没	0.70277	解放思想	0.5518	大举进攻	0.5341	各阶	0.5469	尊严	0.4627	苦苦	0.6586
(the whole army is wiped out)		(liberate thoughts)		(launch a large-scale attack)		各阶		(dignity)		(strenuously)	
大举进攻	0.7011	正确	0.5435	危急	0.4671	充满	0.5443	另起炉灶	0.4616	大	0.5647
(launch a large-scale attack)		(correct)		(critical)		(be brimming with)		(make a fresh start)	0.4616	(big)	0.5647
正面	0.6918	准确	0.4973	不均	0.4570	广泛	0.5207	准确	0.4564	坚定	0.5517
(positive)		(Accurate)		(unequal)		(extensive)		(Accurate)		(firm)	
南大	0.6819	明确	0.4922	稳定	0.4506	拨乱反正	0.5142	红 (red)	0.4522	焕然一新	0.5111
(South Big)		(clear-cut)		(stable)		(bring order out of chaos)	0.5142	(red)	0.4522	(take on an entirely new look)	0.5111
壮烈	0.6714	最重	0.4600	恐慌	0.4490	团结	0.4984	满	0.4484	深远影响	0.5019
(heroic)		(heaviest)		(frightened)		(united)		(full)		(have far-reaching influence)	
钦差大臣	0.6678	连续	0.4565	取得胜利	0.4404	团结一致	0.4801	深入人心	0.4066	良好	0.4785
(imperial envoy)		(consecutive)		(get victory)		(united)		(strike a deep chord in the hearts of the people)		(fine)	
公开	0.5995	热烈	0.4218	均匀	0.42925	独特	0.4474	旧	0.3868	独特	0.4718
(make public)		(ardent)		(even)		(unique)		(old)		(unique)	
恐怖	0.5886	星星之火	0.4160	残暴	0.3887	焕然一新	0.4465	易	0.3815	精神文明	0.4696
(scary)		(a single spark can start a prairie fire)		(brutal)		(take on an entirely new look)		(simple)		(spiritual civilization)	

Notes: We use Jiena's part-of-speech tagging to define adjectives in this analysis.

Table 5: Hong Kong closest adjectives

Chiang (蒋介石)		Mao (毛泽东)		KMT (国民党)		CCP(共产党)		Sun(孙中山)		Marx(马克思)	
Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance
束手无策	0.7419	老	0.5909	束手无策	0.6012	薄弱	0.5646	薄弱	0.5406	独秀	0.7022
(be at a loss at what to do)		(old)		(be at a loss at what to do)		(weak)	0.5646	(weak)		(duk-sau) (Name)	0.7022
不久	0.7303	乐观	0.5643	重建	0.4973	独秀	0.5627	成功	0.5179	广泛	0.6730
(soon)		(optimistic)		(rebuild)		(duk-sau) (Name)		(success)		(extensive)	
精锐	0.6846	亲密	0.5441	薄弱	0.4798	坚决	0.5195	落实	0.4874	充满	0.6220
(selected)		(close)		(weak)		(firm)		(relaxed)		(be brimming with)	
公开	0.6456	整肃	0.5313	壮大	0.4756	团结	0.5041	著名	0.4414	不良	0.6043
(open)		(enforce)		(strong)		(unite)		(famous)		(bad)	
大势已去	0.6018	最高	0.5296	健康	0.4707	团结一致	0.4662	激愤	0.4253	着重	0.5947
(the game is as good as lost)		(the highest)		(healthy)		(united)		(indignant)		(emphasize)	
烈	0.5929	独秀	0.5183	精锐	0.4685	最高	0.4557	不久	0.4237	封建	0.5920
(intense)		(duk-sau) (Name)		(selected)		(highest)		(soon)		(feudal)	
义勇	0.5599	大中	0.5106	坚决	0.4602	大势已去	0.4441	独秀	0.4128	不适	0.5905
(righteous and courageous)		(central)		(firm)		(the game is as good as lost)		(duk-sau) (Name)		(unsuitable)	
激愤	0.5376	言者无罪	0.5016	强烈不满	0.4560	大党	0.4350	封建	0.3999	优越	0.5870
(indignant)		(blame not the critics)		(strong resent)		(big party)		(feudal)		(superior)	0.5870
速战速决	0.5305	面红	0.4904	各阶	0.4261	精锐	0.4316	长	0.3966	著名	0.5755
(fight a quick battle)		(be red in the face)		every aspect		(selected)		(long)		(famous)	
突然	0.5202	公开	0.4657	持久	0.4247	落实	0.4186	确实	0.3678	偷工减料	0.5656
(suddenly)		(public)		(lasting)		(implement)		(real)		(do shoddy work and use inferior materials)	

Notes: We use Jiena's part-of-speech tagging to define adjectives in this analysis.

Table 6: TW closest adjectives

Chiang (蒋介石)		Mao (毛泽东)		KMT (国民党)		CCP(共产党)		Sun(孙中山)		Marx(马克思)	
Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance	Word	
顺利	0.7834	最高	0.7222	最高	0.7448	最高	0.8217	屡遭	0.6208	旧	0.8652
(smooth)	0.7834	(the highest)		(the highest)	0.7448	(the highest)		(to suffer repeatedly)		(old)	
最高	0.7704	重新	0.7029	重新	0.6985	屡遭	0.7415	友好	0.5914	深为	0.8635
(the highest)		(again)		(again)		(to suffer repeatedly)		(friendly)		(be deeply)	
取得胜利	0.7652	严峻	0.6317	严峻	0.6489	小	0.6811	顺利	0.5526	年轻	0.8239
(to get victory)		(severe)		(severe)		(small)		(smooth)		(young)	
正式	0.6539	整肃	0.5313	成功	0.6337	成功	0.6698	成功	0.5436	优越	0.8165
(formal)		(enforce)		(success)		(success)		(success)		(superior)	
成功	0.6431	崭露头角	0.5907	屡遭	0.6238	稳固	0.6666	有利	0.5375	蓬勃发展	0.8015
(success)		(come to prominence)		(to suffer repeatedly)	0.6238	(stable)		(advantageous)		(develop vigorously)	
诚	0.6141	成功	0.5888	顺利	0.6069	重新	0.6633	大势已去	0.5273	各阶	0.7983
(sincere)		(success)		(smooth)		(again)		(the game is as good as lost)		(all sectors)	
小	0.6070	腐败	0.5518	腐败	0.5700	严峻	0.6408	远	0.5199	自由	0.7967
(small)		(rotten)		(rotten)		(severe)		(far away)		(free)	
友好	0.5991	强烈	0.5417	稳固	0.5522	团结	0.6385	不幸	0.4984	百家争鸣	0.7915
(friendly)		(strong)		(stable)		(united)		(unfortunate)		(contention and flourishing of numerous schools of thought)	
屡遭	0.5891	广泛	0.5402	小	0.5453	顺利	0.6369	弱	0.4935	救亡图存	0.7886
(to suffer repeatedly)		(widespread)		(small)		(smooth)		(weak)		(save the nation from doom and strive for its survival)	
固守	0.5793	激烈	0.5376	灵活	0.5425	有利	0.6286	矛盾	0.4657	伤害	0.7880
(defend tenaciously)		(violent)		(flexible)		(advantageous)		(contradiction)		(harm)	

Notes: We use Jiena's part-of-speech tagging to define adjectives in this analysis.

Table 7: Number of Mentions of politically-important figures and entities

Region	Version	Mao	CCP	Marx	Chiang	KMT	Sun
Hong Kong	Manhattan	104	226	10	30	91	63
Hong Kong	Modern	82	331	8	41	151	68
Mainland	Renjiao	63	91	33	10	94	35
Mainland	Renmin	73	147	47	8	110	36
Mainland	Xueli	47	100	19	12	85	21
United States	Search of Modern China	403	1109	91	290	439	201
Taiwan	Nane	25	94	0	21	38	35
Taiwan	Kangxi	25	96	4	21	29	21
Taiwan	Lungtun	44	121	4	17	74	33
Mainland	old	62	152	14	41	116	37
Taiwan	old	2	74	1	42	32	39

Region	Version	1	2	3	4	5
Hong Kong	Manhattan	支持 (support) 57	新 (new) 47	积极 (active) 38	和平 (peace) 35	平等 (fair) 34
Hong Kong	Modern	和平 (peace) 76	支持 (support) 62	积极 (active) 46	统一 (unified) 43	重要 (important) 43
Mainland	Renjiao	新 (new) 93	和平 (peace) 40	科学 (science) 30	重要 (important) 27	统一 (unified) 26
Mainland	Renmin	新 (new) 133	科学 (science) 72	和平 (peace) 56	基础 (foundation) 51	主要 (primary) 47
Mainland	Xueli	新 (new) 75	科学 (science) 47	和平 (peace) 43	重要 (important) 43	统一 (unified) 34
United States	Search of Modern China	新 (new) 347	重要 (important) 179	支持 (support) 164	主要 (primary) 142	接受 (accept) 137
Taiwan	Nane	新 (new) 28	重要 (important) 28	中正 (center) 21	积极 21	主要 (primary) 19
Taiwan	Kangxi	传统 (tradition) 31	重要 (important) 30	支持 (support) 26	主要 (primary) 23	中正 (center) 21
Taiwan	Lungtun	主要 (primary) 42	新 (new) 41	重要 (important) 38	支持 (support) 35	传统 (tradition) 27
Mainland	old	新 (new) 83	主要 (primary) 65	和平 (peace) 57	统一 (unified) 50	基础 (foundation) 32
Taiwan	old	统一 (unified) 32	美 (beautiful) 23	和平 (peace) 22	允 (allow) 20	平等 (fair) 20

Table 8: Top 5 Positive Phrases

The number next to the phrase representing the number of times that phrase shows up. The column number represents the rank of the phrase. The phrase 'Revolution' and 'Economy' are excluded from the list.

Region	Version	1	2	3	4	5
Hong Kong	Manhattan	严重 (serious) 48	反 (oppose) 23	不平 (unfair) 21	专制 (dictatorial) 16	难以 16
Hong Kong	Modern	严重 (serious) 52	反 (oppose) 40	难以 25	官僚 (bureaucracy) 18	不平 (unfair) 15
Mainland	Renjiao	封建 (feudal) 46	斗争 (struggle) 20	严重 (serious) 16	官僚 (bureaucracy) 10	错误 (mistake) 10
Mainland	Renmin	封建 (feudal) 55	斗争 (struggle) 51	严重 (serious) 28	官僚 (bureaucracy) 20	专制 (dictatorial) 11
Mainland	Xueli	斗争 (struggle) 27	严重 (serious) 19	专制 (dictatorial) 16	旧 (old) 15	错误 (mistake) 13
United States	Search of Modern China	似乎 (as if) 141	官僚 122	攻击 (attack) 118	严重 (serious) 99	抗议 (protest) 97
Taiwan	Nane	严重 (serious) 17	反 (oppose) 8	讨 (beg for) 8	名义 (name) 6	所谓 (so-called) 6
Taiwan	Kangxi	斗争 (struggle) 17	策略 14	严重 (serious) 13	所谓 (so-called) 10	专制 (dictatorial) 9
Taiwan	Lungtun	反 (oppose) 19	斗争 (struggle) 17	严重 (serious) 16	动荡 (unstable) 11	官僚 (bureaucracy) 11
Mainland	old	封建 (feudal) 87	斗争 (struggle) 72	严重 (serious) 42	错误 (mistake) 37	反 (oppose) 23
Taiwan	old	复 (turn over) 27	所谓 (so-called) 16	严重 (serious) 13	不平 (unfair) 11	伪 (fake) 9

Table 9: Top 5 Negative Phrases

The number next to the phrase representing the number of times that phrase shows up. The column number represents the rank of the phrase.

Table 10: Positive and Negative Words in Each Version

Region	Version	Positive	Negative	Positive to Negative Ratio
Hong Kong	Manhattan	3376	1003	3.366
Hong Kong	Modern	4088	1149	3.558
Mainland	Renjiao	2415	531	4.548
Mainland	Renmin	3721	740	5.028
Mainland	Xueli	2404	495	4.857
United States	Search of Modern China	21171	8367	2.530
Taiwan	Nane	1605	581	2.762
Taiwan	Kangxi	1433	465	3.082
Taiwan	Lungtun	1973	564	3.498
Mainland	old	3217	833	3.862
Taiwan	old	2002	636	3.148

Table 11: positive to negative ratio of top 20 closest adjectives

	Chiang	Mao	KMT	CCP	Sun	Marx
Mainland	1.0	5.0	0.5	3.5	4.5	6.0
Hong Kong	2.5	3.0	3.5	3.0	1.25	0.6
Taiwan	6.0	1.0	3.0	2.25	1.5	2.0

Table 12: Adjective and positive-to-negative ratio in major newspapers of the 3 regions

	Oriental Daily Hong Kong	Shanghai Morning Post Mainland	Liberty Times Taiwan
Adjective	168272	98368	50709
Word Count	1403410	777956	389745
Adjective Ratio	0.120	0.126	0.130
Positive	59298	36255	17595
Negative	16988	8663	4096
Positive-Negative Ratio	3.491	4.185	4.296

Figure 3: Pro-CCP to KMT entities

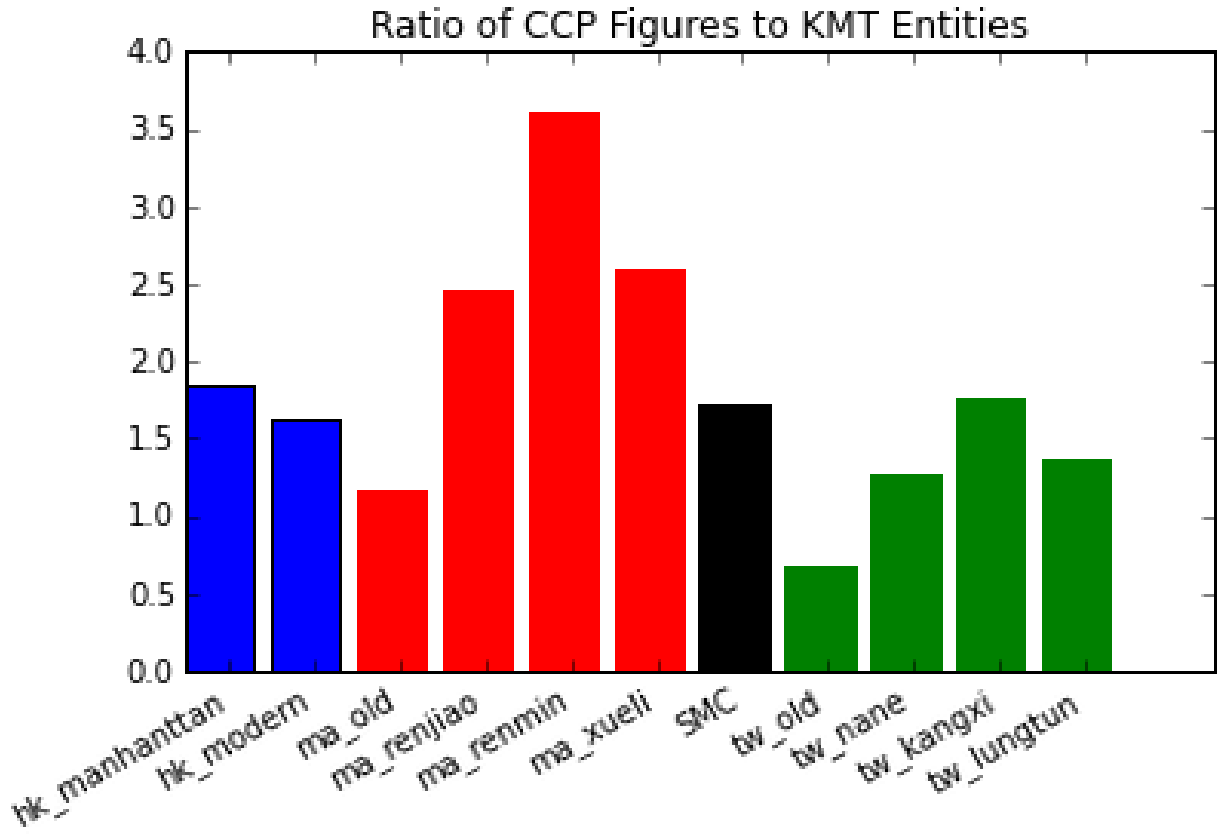


Figure 4: Adjective Ratio

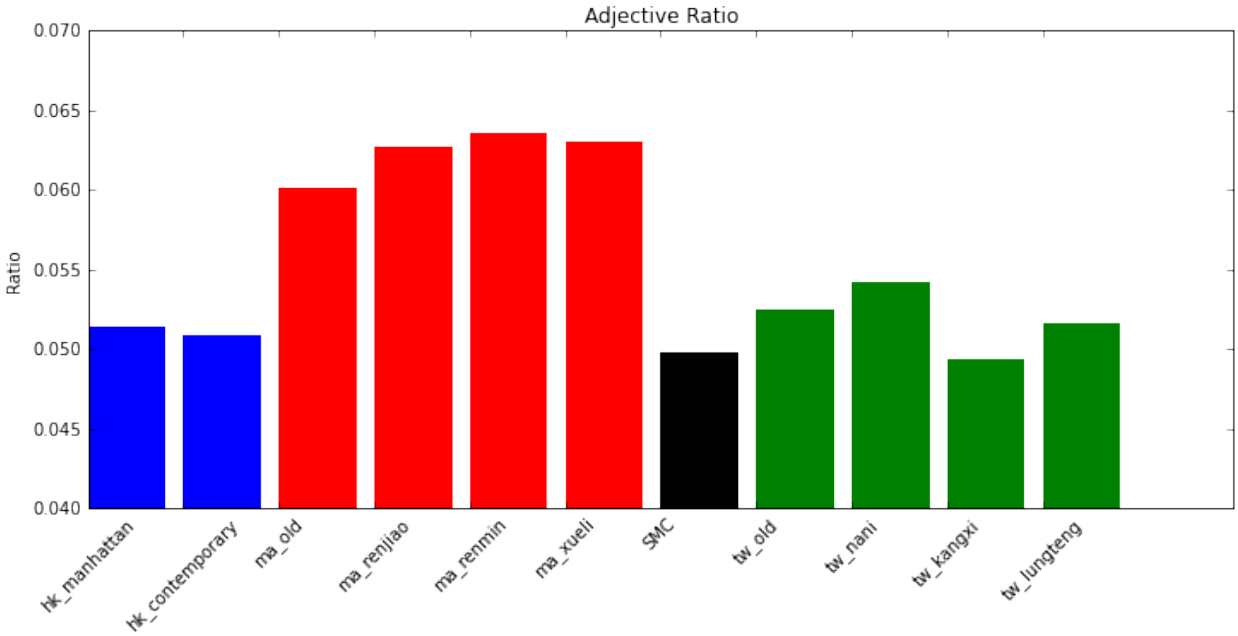


Figure 5: Adverb Ratio

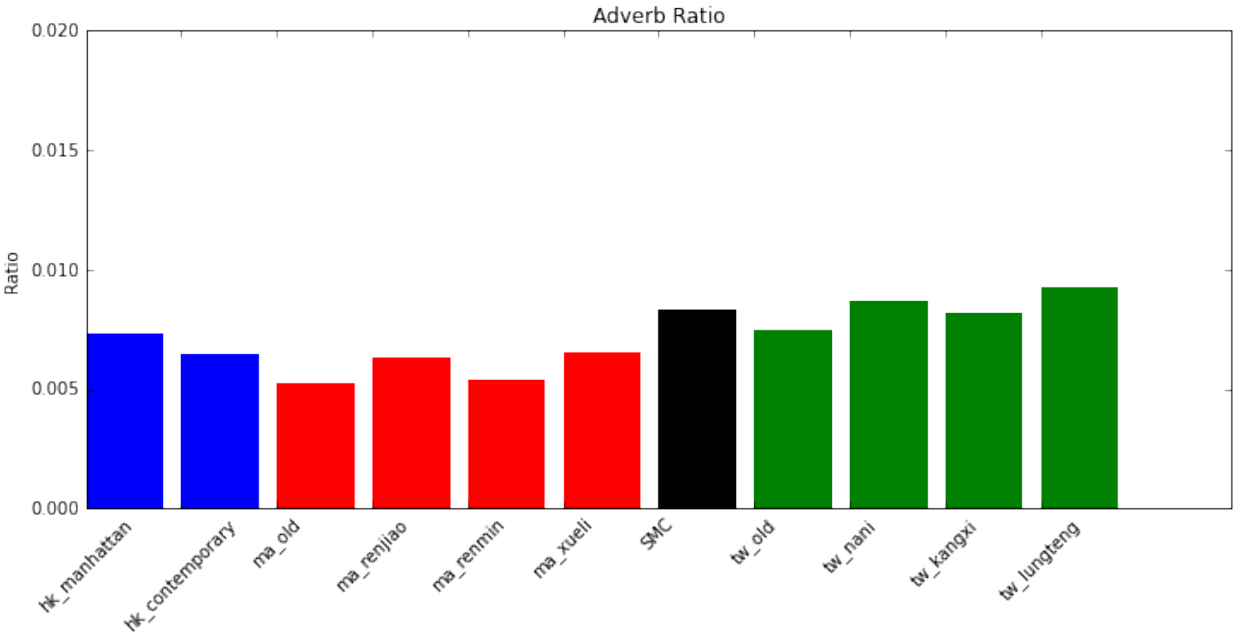


Figure 6: Adjective Ratio of pre-1949 events

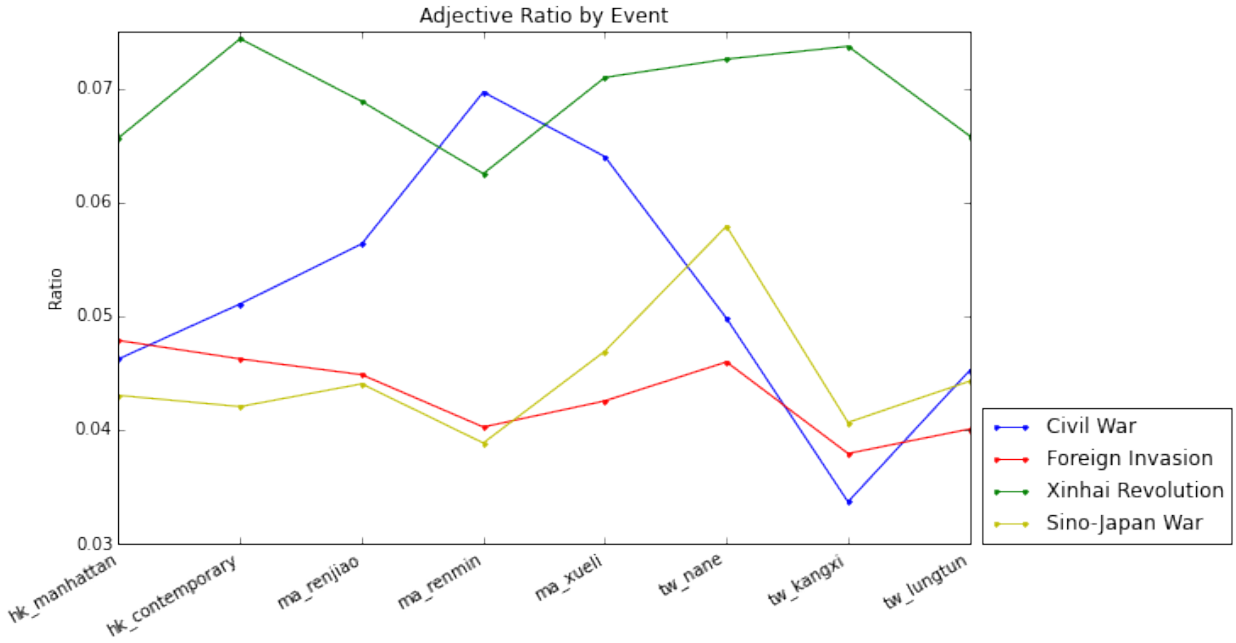


Figure 7: Adjective Ratio of post-1949 events

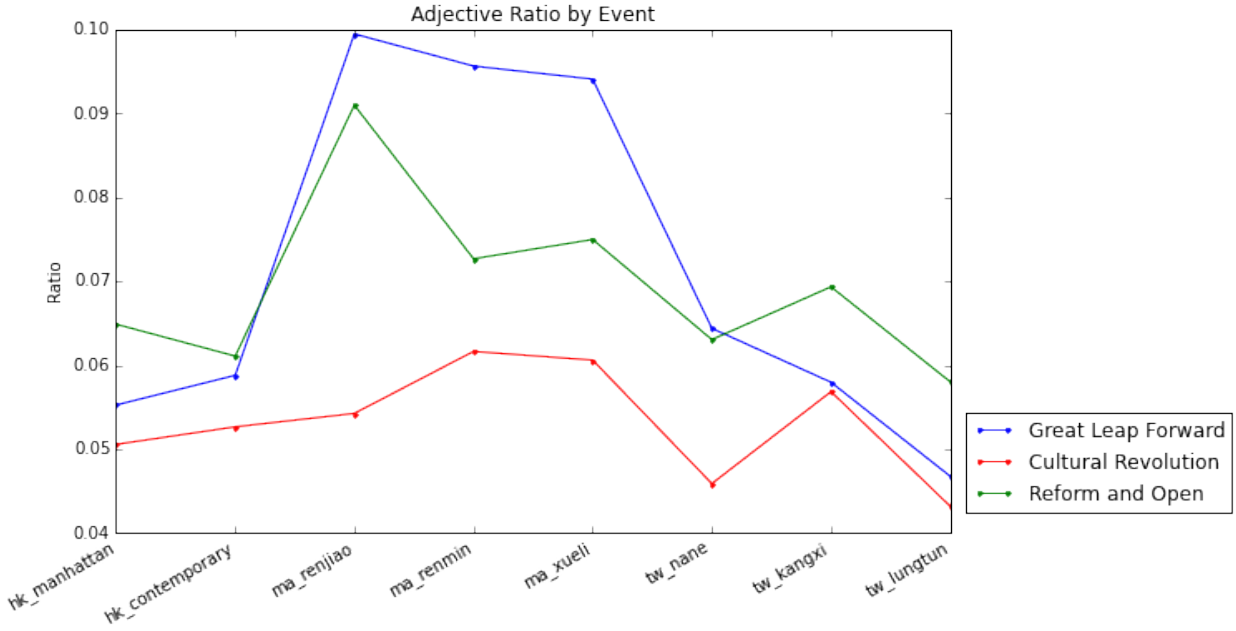


Figure 8: Adjective Ratio pre- and post- 1949

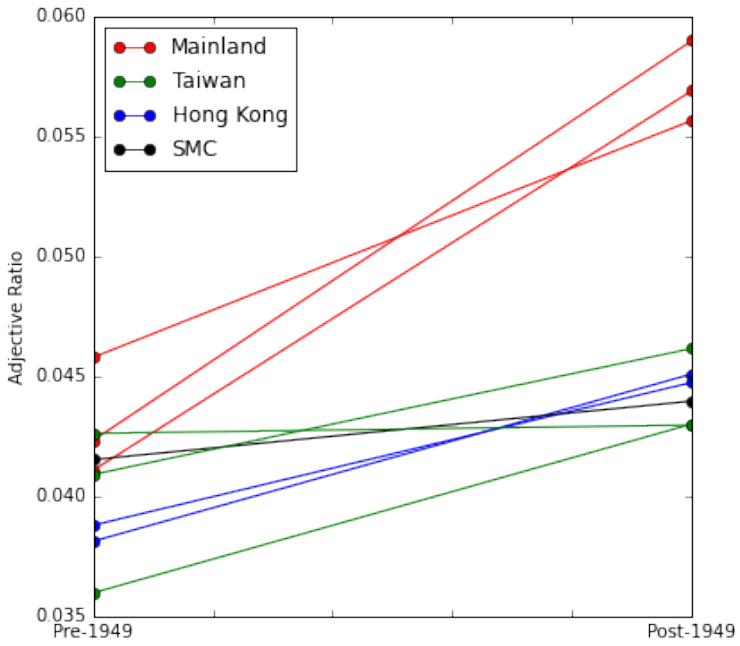


Figure 9: Positive to Negative Phrase Ratio pre- and post- 1949

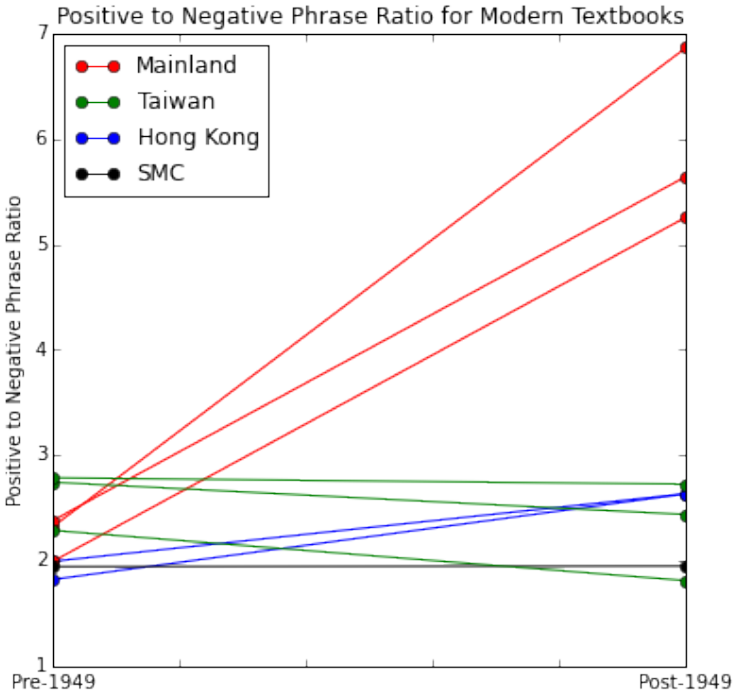


Figure 10: Positive to Negative Phrase Ratio in 4 major pre-1949 Historical Episodes

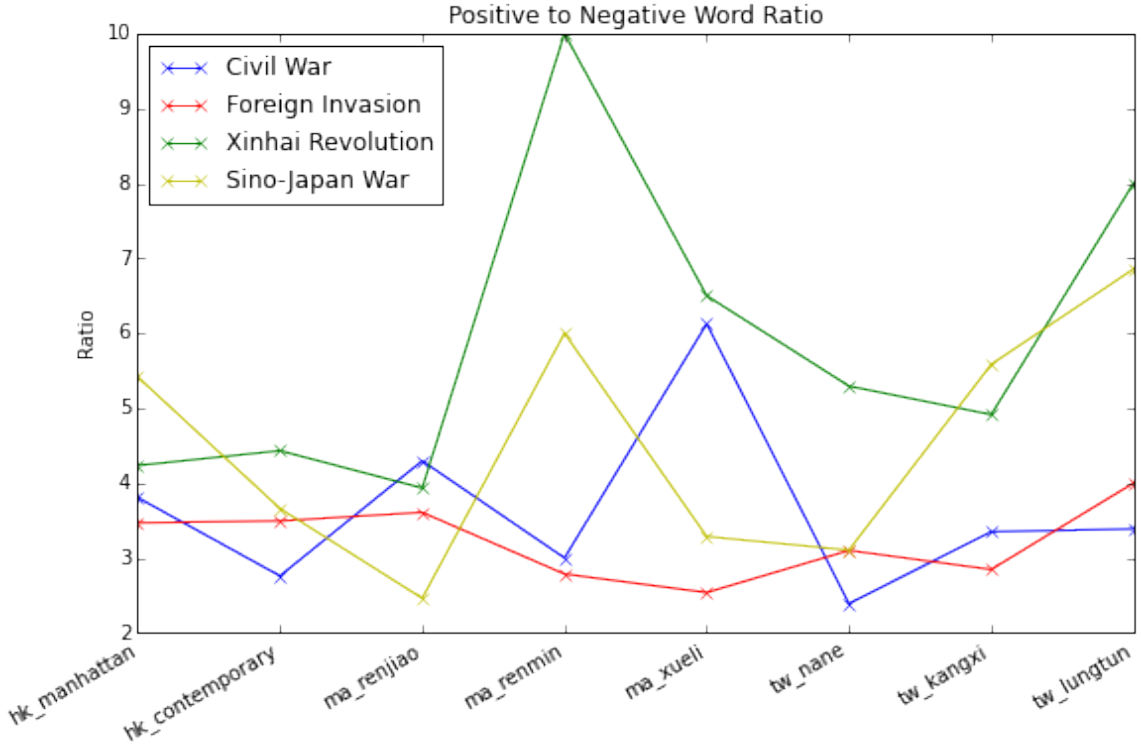


Figure 11: Positive to Negative Phrase Ratio pre- and post- 1949

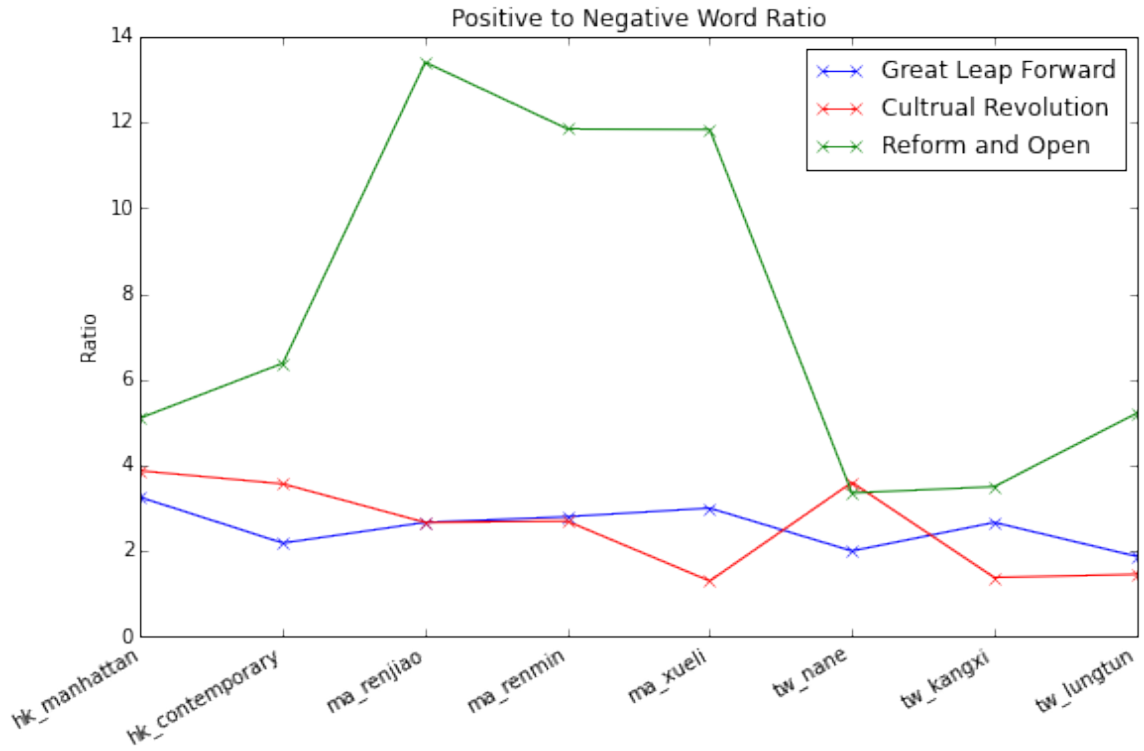


Figure 12: Adjective Ratio in Old and New Versions on Individual Events

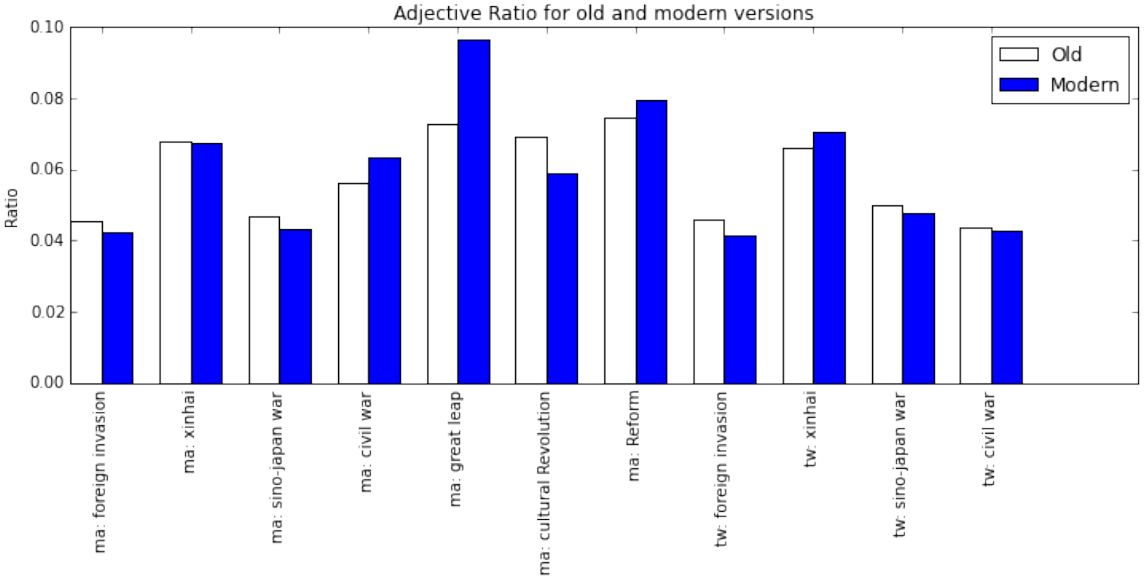


Figure 13: Adjective Ratio in Old and New Versions in pre- and post-1949

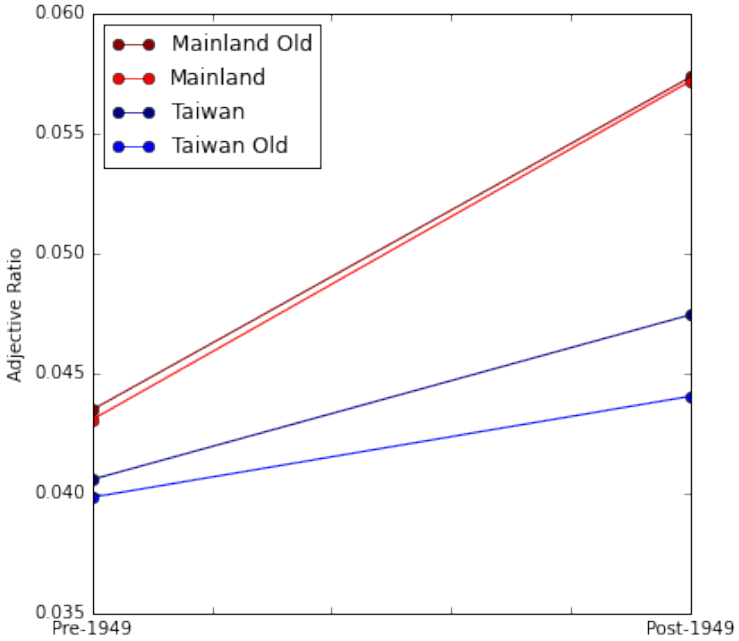
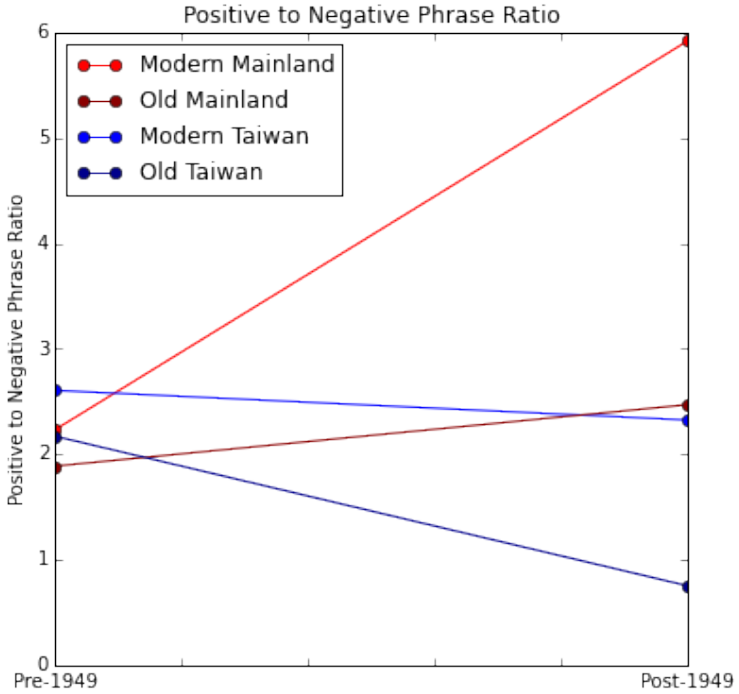


Figure 14: Positive to Negative Phrase Ratio pre- and post- 1949

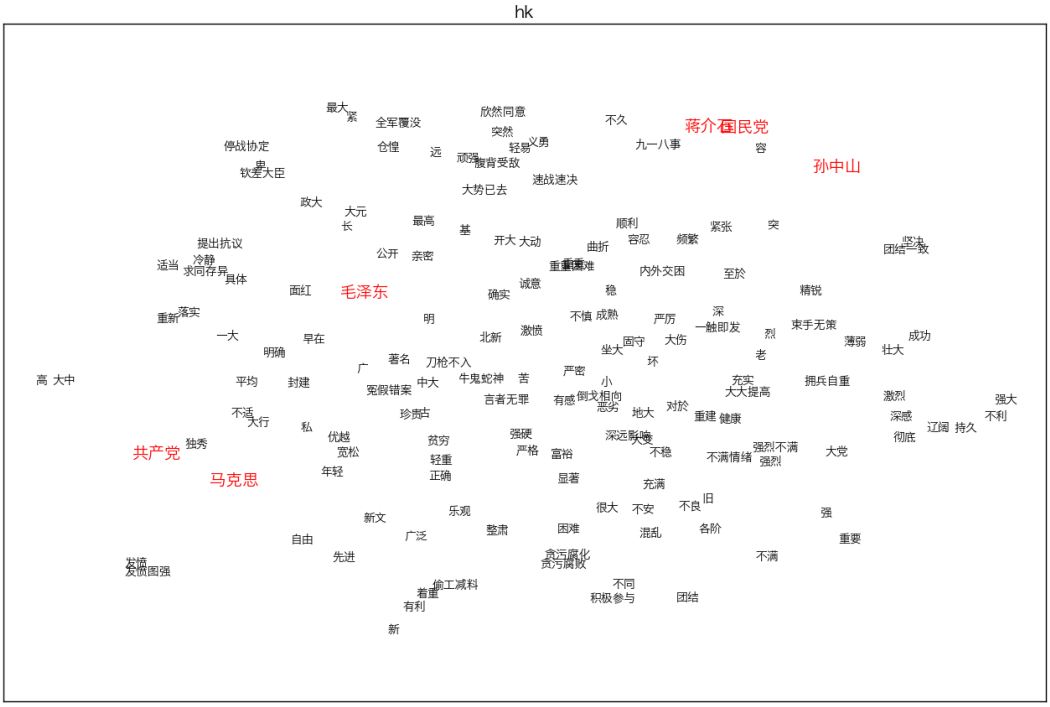


Note: The adjective ratio for modern mainland versions is calculated using all 3 books

Figure 15: word embedding visualization of Mainland Chinese textbooks by t-SNE

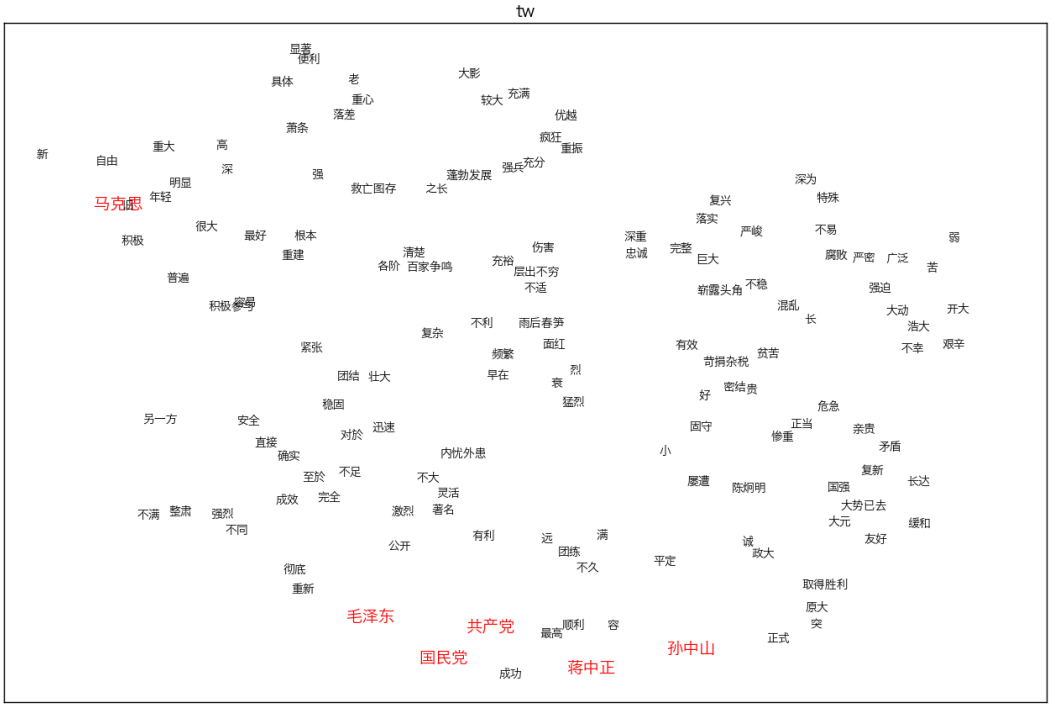
Note: The phrases highlighted in red are the 6 political entities of interest: 蒋介石 (Chiang Kai-Shek), 毛泽东 (Mao Ze-dong), 孙中山 (Sun Yat-sen), 国民党 (Kuomintang), 共产党 (Chinese Communist Party), 马克思 (Marx)

Figure 16: word embedding visualization of Hong Kong textbooks by t-SNE



Note: The phrases highlighted in red are the 6 political entities of interest: 蒋介石 (Chiang Kai-Shek), 毛泽东 (Mao Ze-dong), 孙中山 (Sun Yat-sen), 国民党 (Kuomintang), 共产党 (Chinese Communist Party), 马克思 (Marx)

Figure 17: word embedding visualization of Taiwanese textbooks by t-SNE



Note: The phrases highlighted in red are the 6 political entities of interest: 蒋介石 (Chiang Kai-Shek), 毛泽东 (Mao Ze-dong), 孙中山 (Sun Yat-sen), 国民党 (Kuomintang), 共产党 (Chinese Communist Party), 马克思 (Marx)