A Computational Analysis of History Textbooks in Mainland China, Hong Kong and Taiwan

Onyi Lam[1] and Eddie Lin[2]

Abstract— **This study analyzes the linguistic features of high school history textbooks used in mainland China, Hong Kong, and Taiwan. Mainland history textbooks are found to use more quadgrams, have a higher adjective ratio overall, and a higher positive-to-negative phrase ratio in post-1949 content. Moreover, the ratio of Chinese Communist Party (CCP) entities to Kuomintang (KMT) entities is higher in mainland textbooks. By studying the textbooks before and after the curriculum reform, we also found that the positive-to-negative phrase ratio in post-1949 history content has registered a considerable increase over time in mainland textbooks.**

I. Introduction

History textbooks provide a lens through which students view the nation's past. By conveying a uniform and approved narrative of the nation's past, history textbooks help instill a shared identity in youth. This shared interpretation of history is essential to nation-building and can shape contemporary patriotism. In this context, it is often in the interest of authorities, especially that of restrictive regimes, to use history textbooks as an "ideological tool to promote certain beliefs that legitimize a political order" (Apple and Christian-Smith, 1991).

Compared to other information channels such as social media or newspapers, history textbooks could be easier for the authority to manipulate. Unlike the mass media, history textbooks are mandatory in the curriculum for most students. Students are also highly incentivized to learn the material in order to do well in exams and read them at a young age. Studies have shown that individual values and beliefs are more susceptible to outside influence at a younger age.[1]

The present study analyzes the high school history textbooks used in mainland China, Hong Kong, and Taiwan. While both Hong Kong and Taiwan share historical roots with mainland China, both have political systems that are separated from the Chinese mainland for long periods. Young people in the three regions have also developed very different attitudes toward the Chinese Communist regime and the Chinese identity. A telling example is the recent protest in Hong Kong, which demanded more democratic local institutions. While Taiwanese people and the government have expressed sympathy for the movement, the protest has mostly met with scorns from the mainland. And even the highly educated mainland Chinese youth express strong approval for the Chinese government and frequently hold counter-protests in the pro-Democracy rallies that support Hong Kong protesters on overseas university campuses. Wang (2008) noted that one reason for this "patriotism" is that history education has been successfully used by Beijing "as an instrument for the glorification of the party," and the narrative that the CCP has led China to overcome "one hundred years of humiliation" has been bolstered by China's strong economic growth over the last two decades.

Using tools and insights developed in the field of natural language processing, our empirical analysis aims to further explore the role of history education in CCP's attempt to justify the regime's legitimacy and shed light on the following questions: Are there detectable differences in emphasis and writing style across the history textbooks used in the three regions? Can we associate certain linguistic characteristics with textbooks of a particular region? Do mainland history textbooks use more subjective and emotional linguistic features that are more effective in influencing students' political attitude? The answers to these questions can provide concrete evidence to the claim that history education is being used to justify "the political system of the CCP's one party rule." For this purpose, we specifically emphasize on the metrics that capture the subjective elements in the text.

II. Related Work

This study is most related to the literature that concerns the use of natural language processing techniques to elicit political content in various communication channels (for example: Gentzkow and Shapiro 2010; Tumasjan et al. 2010; Beauchamp 2016; Roberts et al. 2016; Savoy 2010, 2016; Theocharis et al. 2017). These data-intensive techniques are often suitable in analyzing a large amount of text and allow researchers to score the ideology and assign similarity metrics to lawmakers and media by the language they use. While much of the focus is on political speech as well as political discourse on social media and newspapers, our focus on education material provides new findings in a varied setting.

*This work was not supported by any organization

[1] FiscalNote. This work was done while the author was at University of California, San Diego

[2] University of Chicago

[1]Yolanda Xue provided excellent research assistance.

[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.

Previous studies on textbooks in East Asia (Cantoni et al. 2016; Shin and Sneider 2011; Ye 2016) have also explored their political implications. Cantoni et al. (2016) in particular documents evidence of the persuasive effect of textbooks on students' political attitude using a carefully designed causal identification strategy that exploits the differential introduction date of a new curriculum across provinces in China. Shin and Sneider (2011) examine how the description of World War II differs across history textbooks used in China, Japan, and Korea. Ye (2016) analyzed how the textbooks differ between mainland China and Taiwan in a more general sense. Much of this literature relies on human interpretation which inevitably involves a certain degree of subjective judgment. The quantitative approach used in this study provides new ways to analyze and quantify the content and can help mitigate the potential human bias in interpretation. Our analysis also included textbooks that were used before curriculum reform to shed light on how textbooks changed over time. Outside of East Asia, researchers have found that significant differences exist between the Israeli and Palestinian textbooks in which what was positive on one side could be negative on the other side (Jala, 2003). These differences were said to breed hatred and contribute to ongoing conflicts.[2]

III. History Textbooks in Mainland China, Hong Kong and Taiwan

Our textbook samples, which were in use as of 2016, consist of three publisher versions from the Mainland, two from Hong Kong, and three from Taiwan. They are some of the most widely used versions in their respective regions. Students exposed to these textbooks are generally between the ages of 15 to 17. We purchased the physical copies through an online shopping platform. An undergraduate research assistant then digitalized the textbooks into searchable text files by using both voice typing and manual entry. We only use the content after the Opium War (A.D. 1842) and ignore all text in captions and appendices.

The three mainland publisher versions are all officially approved: Renmin, Renjiao, and Yuelu. The publishers are either state-owned or subsidiaries of state-owned enterprise. Provinces and cities can choose one of the three approved versions for their high schools.[3] Renjiao is the most popular of the three versions with the largest number of adopting regions. Despite differences in adoption across provinces, the organization of the curriculum is very similar across the versions: each version comprises of three books and each book focuses on a specific aspect. The first book presents the historical events in a generally chronological order. The second book focuses on the economic development, whereas the third book focuses on cultural and technological progresses in the corresponding time period. Information not directly related to China is also presented sometimes, e.g., the different political systems used in other parts of the world as well as important western literature and scientific advancements of the concurrent time period. We exclude these discussions in our text analysis to ensure content is comparable across regions. In addition, the curriculum is separated into a mandatory and an elective portion, but we use the mandatory portion only.

In both Hong Kong and Taiwan, the respective Education Bureau issues guidance for private publishers, who are then responsible for writing the textbooks. The schools have the discretion to decide which publisher's version to use. In Taiwan, there are seven popular versions in the market, and our sample consists of three of these: Kangzi, Nane, and Lungtun. Kangzi is the most popular textbook and is used by 45% of the schools, followed by Nane (17%) and Lungtun (16%).[4]. The curriculum is organized into four parts with respective focus on Taiwan History, Chinese History, Ancient World History, and Modern World History.[5] To ensure comparability across regions, we use the part on Chinese History only. The two Hong Kong publisher versions are Manhattan Press and Modern Educational Research Society Limited. [6] Manhattan Press separates the whole curriculum into six books while Modern Education has four books.

History textbooks in both the Mainland and Taiwan have undergone significant revisions in the last two decades. We include in our analysis two old textbooks that were used prior to curriculum reforms in mainland China and Taiwan. The old Taiwan history textbook was used between 1983 and 1999 before the textbook market opened up for private publishers. Prior to that, history textbook content was standardized by the Education Bureau. On the other hand, the old mainland Chinese textbook was used in 2003, before China's 8th Curriculum Reform took place in 2004. The Chinese government had launched the "Patriotic Education Campaign" in the early 1990s after the 1989 Tiananmen Movement, and the 2004 curriculum reform has an explicit goal to instill "a correct worldview" in students (Cantoni et al., 2015). We also obtained a translated version of a Chinese History textbook commonly used in the introductory classes by universities in the United States: "The Search for Modern China" (SMC). SMC is written by British-American scholar Jonathan Spence and is relatively devoid of influences of a state political agenda. It provides an interesting reference point for

[2] http://www.economist.com/blogs/pomegranate/2013/02/israeli-and-palestinian-textbooks

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

[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 publisher 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

the linguistic features considered in this study. Granted, there are important differences between SMC and the high school textbooks used in the three regions. Specifically, the scope of SMC is wider and it is significantly longer than any of the history textbooks. SMC also provides more information about the cultural and social context by introducing and describing in detail the lives of people who are lesser known in history, whereas the narrative of the history textbooks is more consistent and unambiguous.

IV. Analysis

As part of preprocessing, we converted the text into simplified Chinese and removed all numbers, punctuation, and non-Chinese characters. Unlike English, there is no white space between words in Chinese, so we had to segment the text into meaningful phrases. To do that, we used Jieba, a Python Chinese word segmentation module. We also supplemented a list of individual names to Jieba dictionary to help it distinguish individual names from other parts of the sentence. After segmenting the sentences, we removed the stop words from the text.[7]

Table I provides some descriptive statistics of each publisher version. The first column presents the total character count, which varies considerably across regions but not as much within. Textbooks in Hong Kong are the longest on average, followed by that of the Mainland and Taiwan. Columns (2) to (4) present the number of bigrams, trigrams, and quadgrams that Jieba detected. In all versions, the number of trigrams and quadgrams is less than the number of bigrams. In addition, the number of quadgrams exceeds the number of trigrams in all the Mainland versions while the reverse holds true for the Hong Kong and Taiwanese versions. Quadgrams are often superlative adjectives and four-character idioms, which contain a fable with a practical lesson. [8]

Columns (5) and (6) show the character count of pre- and post-1949 content. In all textbooks, the character count of post-1949 history is lesser than that of pre-1949. On average, the fraction of post-1949 character count is the lowest in Taiwan textbooks and highest in the Mainland textbooks, pointing to a stronger emphasis on post-1949 history in the Mainland textbooks. On the other hand, the fraction of post-1949 content is similar between Hong Kong textbooks and the SMC. The bottom two rows of the table provide the summary statistics of the pre- and post-reform textbooks. Both the Mainland and Taiwan pre-reform textbooks have a greater word count than all of their post-reform counterparts and allocate a larger fraction of its total word count to pre-1949 events.

A. *Major Historical Events*

We isolate several major pre- and post-1949 historical events and analyze their linguistic characteristics[9]. Events such as the Cultural Revolution, which is widely acknowledged as the result of a policy blunder, are expected to receive a larger differential emphasis across regions, whereas events that are not directly related to the perception of the regime such as the Xinhai Revolution should have received similar emphasis.

The top panel in table II shows the number of characters dedicated to the four major historical events prior to 1949: Foreign Invasion (1842-1911), Xinhai Revolution (1911), Second Sino-Japanese War (1937-45), and the Chinese Civil War (1945-1949). Foreign Invasion refers to all the wars involving a foreign power since the Opium War but excluding the second Sino-Japanese War. Xinhai Revolution includes a description on Sun Yat-sen and all immediate events leading to the Revolution that ended the Qing Dynasty. 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 surrounding the conflict between CCP and Kuomintang (KMT) after the surrender of Japan in 1945.[10]

Foreign Invasion has the largest character count and ratio among the pre-1949 events in both Hong Kong versions. Despite the notion that the Mainland textbooks like to emphasize the suffering inflicted by foreign powers, the character count on Foreign Invasion and the Sino-Japanese War does not appear to be very different from the Taiwanese textbooks. In contrast, both the pre-reform Mainland and pre-reform Taiwanese textbook dedicate a larger fraction of the overall word count to these two historical episodes. The word count also varies significantly even across versions within the same region. For example, Mainland's Yuelu only has 618 characters on the Sino-Japanese war, whereas Renmin has 3293 characters. By relative length, the Civil War receives the heaviest emphasis in the three Taiwan textbooks.

The bottom panel of table II shows the number of characters dedicated to the three post-1949 historical episodes: Great Leap Forward (1958-61), Cultural Revolution (1966-76), and the Reform and Open Policy (1976 – present, encompassing the Tiananmen Square Protests in 1989). In all versions, the Reform and Open Policy has the most characters. Both the Great Leap and the Cultural Revolution receive smaller coverage in the Mainland textbooks, but the pre-reform Mainland textbook has a noticeably higher ratio on Cultural Revolution. On the other hand, the Cultural Revolution has the highest relative length in the two Hong Kong versions. Interestingly, on the coverage

[7] The list of Chinese stop words comes from an online repository: https://gist.github.com/dreampuf/5548203

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

[9] Since Mainland versions have different emphases in each book, we exclude the portion on economics and culture and technology, and only use the text in the general history.

[10] We do not include text from SMC in the event-level analysis because its writing style, which tends to intervene with multiple themes, makes it difficult to single out the portion that is relevant to the specific historical episode.

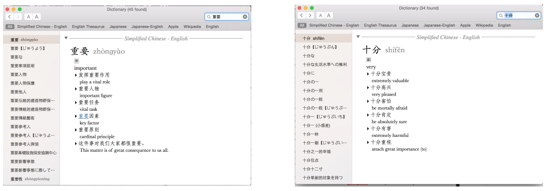
TABLE I: Summary Statistics of Each Publisher Version

Region	Book	Book length	Bigram	Trigram	Quadgram	Pre-1949	Post-1949
Hong Kong	Manhattan	70627	26206	1317	1070	42466	28161
			0.37	0.02	0.02	0.60	0.40
Hong Kong	Modern	91085	33866	1502	1356	58114	32971
			0.37	0.02	0.01	0.64	0.36
Mainland	Renjiao	41544	16778	867	986	19651	21893
			0.40	0.02	0.02	0.47	0.53
Mainland	Renmin	55479	22433	1226	1469	29325	26154
			0.40	0.02	0.03	0.53	0.47
Mainland	Yuelu	38908	15688	857	941	20029	18879
			0.40	0.02	0.02	0.51	0.49
Taiwan	Nanyi	30792	11112	444	345	23136	7656
			0.36	0.01	0.01	0.75	0.25
Taiwan	Kangxi	33640	12679	506	419	24180	9460
			0.38	0.02	0.01	0.72	0.28
Taiwan	Lungtun	40457	15527	610	540	28606	11851
			0.38	0.02	0.01	0.71	0.29
United States	Search for Modern China	410198	140591	7345	3962	235946	174252
			0.34	0.02	0.01	0.58	0.42
Mainland	old	56925	22746	1357	1416	40397	16528
			0.40	0.02	0.02	0.71	0.29
Taiwan	old	46939	16025	552	347	46665	274
			0.34	0.01	0.01	0.99	0.01

The second row of each version represents the ratio as normalized by book length.

of Reform and Open policy, Lungtun has the highest and Nane the lowest ratio, suggesting considerable variation in emphasis within Taiwanese versions. The Tiananmen Square Protest is noticeably absent in all Mainland versions but is mentioned in all Hong Kong and Taiwanese versions.

Fig. 1: Screenshot of the dictionary application



重要 zhòngyào
重要【じゅうよう】
重要な
重要事項説明
重要人物
重要人物保護
重要他人
重要伝統的建造物群保…
重要傳統的建造物群保…
重要傳統藝術
重要参考人
重要参考人【じゅうよ…
重要参考人探偵
重要基礎設施保安協調中心
重要影響事態
重要影響事態に際して…
重要性 zhòngyàoxing

Simplified Chinese - English

重要 zhòngyào
形
important
发挥重要作用
play a vital role
重要人物
important figure
重要任务
vital task
重要因素
key factor
重要原则
cardinal principle
这件事对我们大家都很重要。
This matter is of great consequence to us all.

(a) the word ’Important’

十分 shífēn
十分【じゅうぶん】
十分な
十分な生活水準への権利
十分に
十分の一
十分の一刑
十分の一税
十分の一税【じゅうぶ…
十分一【じゅうぶいち】
十分一 (小惑星)
十分一銖
十分一銀【じゅうぶい…
十分之一的幸福
十分位点
十分十二寸
十分単射的対象を持つ

Simplified Chinese - English

十分 shífēn
副
very
十分宝贵
extremely valuable
十分高兴
very pleased
十分害怕
be mortally afraid
十分肯定
be absolutely sure
十分有害
extremely harmful
十分重视
attach great importance (to)

(b) the word ’Very’

B. Adjective Ratio

We specifically examine how often adjectives and four-character idioms (Chengyu) are present in the text. The intuition is that provocative and inflammatory language is often aided by the use of adjectives. Adjectives are usually difficult to verify quantitatively (what makes a war "brutal"?), which also makes it difficult for the information receiver to dispute with the information sender. The sentiment analysis literature has also found that the presence of adjectives is strongly predictive and is useful in detecting whether an article belongs to an opinion piece as opposed to a news piece (Hatzivassiloglou and Wiebe 1999; Bruce et al. 1999; Bruce and Wiebe 2000).

Instead of using a part-of-speech (POS) tagger commonly available in text processing software library to detect presence of adjectives, we wrote a script to look up the word definition in the pre-installed Chinese dictionary on a Mac laptop. The reason is that off-the-shelf Chinese POS taggers often miss important descriptive elements in the noun phrase that are similar to adjectives. We therefore analyzed each possible word combination in a phrase to determine if any of them has an adjective component.[11]

A phrase is considered as an adjective if the word definition contains the character 形 (English: adjective) (See figure 1a) and a four-character idiom if the character 成 is present. In addition to adjectives, we also examined the presence of adverbs. A phrase is considered as an adverb if the character 副 is present in the word definition (See figure 1b). Table III lists the top three quadgram adjectives and their respective frequencies in each publisher version. 独立自主 (Act independently and of one’s own initiative) is the most common quadgram for all three Mainland versions. It is worth noting that all popular quadgrams in Mainland versions have a positive connotation. On the other hand, 内忧外患 (Domestic trouble and foreign invasion) is one of the top three quadgrams in both Hong Kong versions.

Figure 2 and 3 plot the adjective and adverb ratio by publisher version. The three Mainland versions have notably higher adjective ratio than the other versions. In comparison, SMC has an adjective ratio similar to Hong Kong and Taiwanese versions. The adverb ratio is slightly higher in Taiwanese versions relative to Hong Kong and Mainland versions but is otherwise similar to SMC. The variation in adverb ratio is also smaller than that of the adjective ratio, with the difference between the largest and the smallest ratio being about 0.7%. Figure 2 also shows that the adjective ratio is marginally smaller in the pre-reform Mainland textbook, while the pre-reform Taiwan textbook has

[11] As an example, 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 (Tsinghua University) is inevitably counted twice because of the different possible combinations within the phrase. However, this allows us to separately examine the individual words within the phrase.

an adjective ratio similar to the modern versions.

Fig. 2: Adjective Ratio

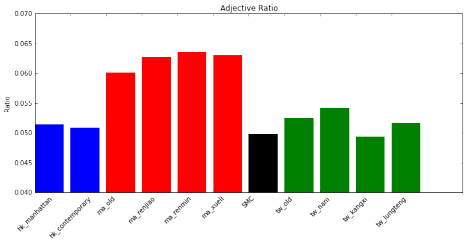


Fig. 3: Adverb Ratio

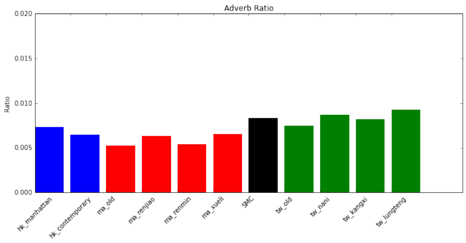


Figure 4 and 5 plot the adjective ratio of the four pre-1949, and the three post-1949 events. There is no visible regional pattern in the pre-1949 events but among the post-1949 events, Great Leap Forward has a noticeably higher adjective ratio in the Mainland versions.

Fig. 4: Adjective Ratio of pre-1949 events

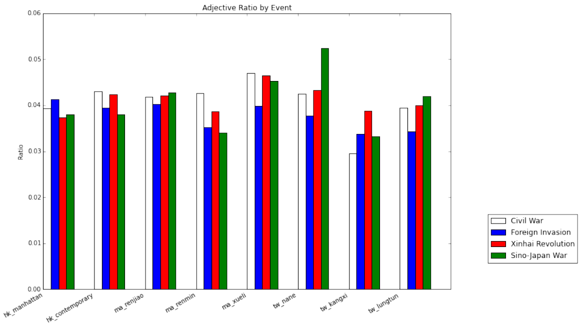


Fig. 5: Adjective Ratio of post-1949 events

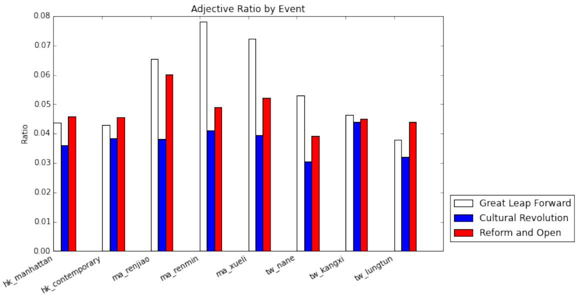
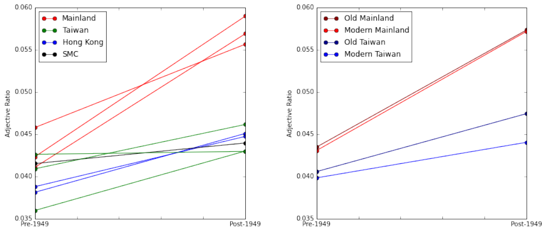


Figure 6 plots the adjective ratio by pre- and post-1949 content. The left-hand side shows the post-reform individual textbooks, and the right-hand side the pre-reform ones along with the average of the modern versions. The pre-1949 adjective ratio is on average the highest in Mainland, even though it is similar to that of SMC and Hong Kong. But strikingly, the average increase from pre- to post-1949 is much larger in Mainland textbooks. On the right-hand side, the increase in adjective ratio from pre- to post-1949 in the old Taiwan textbook is slightly larger than that of the post-reform ones. An increase of a larger magnitude from pre- to post-1949 is seen in the pre-reform Mainland textbook, but the slope is almost identical to that of the post-reform ones.

Fig. 6: Adjective Ratio in Pre- and Post-1949



Note: The left-hand side plots the individual publisher version, and the right-hand side plots the pre-reform Mainland and Taiwan textbooks along with the average of the modern versions

TABLE II: Character Count of Major Historical Episodes by Publisher Version

	Hong Kong		Mainland				Taiwan		Mainland	Taiwan
	Manhattan	Modern	Renjiao	Renmin	Yuelu	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-Japanese 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	-

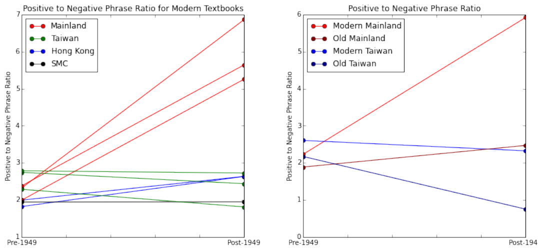
The second row indicates the ratio of the event word count to the book word count. Since the old Taiwan textbook contains a very short description of post-1949 history, we do not separate them into individual historical episodes.

C. Text Polarity

Whether the phrases used to describe a person or an event are positive, negative, or neutral can inform the readers a lot about the author's view on that entity. The binary classification of words is a crude form of sentiment analysis but in absence of a reliable sentiment dictionary that captures more granular emotion classification, the frequencies of positive and negative phrases give a reasonable assessment of the writer's opinion on a given topic. Relatedly, studies in Neuroscience as well as Political Communication have found that text with strong valence tends to be more memorable than neutral adjectives even though the effect of positive and negative messages is asymmetric.[12]

The positive-to-negative phrase ratio allows us to examine a more specific hypothesis that the ruling regime has an incentive to emphasize policy success and underplay the mistakes. We expected the positive-to-negative ratio of specific historical episodes and time periods to vary. And to determine the phrase polarity, we use a pre-existing Chinese sentiment dictionary, which consists of 4570 positive and 4374 negative phrases.[13]

Fig. 7: Positive-to-Negative Ratio in Pre- and Post-1949



Note: The left-hand side plots the individual publisher version, and the right-hand side plots the pre-reform Mainland and Taiwan textbooks along with the average of the modern versions

Table III reports the ratio of positive to negative phrases by version. While all the versions have a ratio of more than 1; Mainland versions have the highest ratio of all, followed by Hong Kong and Taiwan. In particular, Renmin's ratio is approximately twice the lowest ratio version - SMC.[14]

The left hand side of figure 7 plots the pre- and post-1949 positive-to-negative phrase ratio for the individual publisher versions. In the pre-1949 period, the ratio is less than 3 for all the versions but Mainland versions record the largest increase from pre- to post- 1949 with a post-1949 value between 5 to 7 positive to 1 negative phrase. In contrast, Hong Kong versions have a much smaller increase, while Taiwanese versions record a mild decrease on average. SMC has roughly similar ratio in pre- and post-1949. The right hand side of figure 7 plots the same ratio for pre-reform textbooks and the average of the modern versions. Compared to the pre-reform version, the post-reform Mainland versions have a much larger increase in positive-to-negative ratio in the post-1949 content, whereas the opposite holds for Taiwanese textbooks. This pattern is consistent with the notion that there is more emphasis on the positive aspects of post-1949 history in the Mainland textbooks.

Fig. 8: Positive to Negative Phrase Ratio in 4 major pre-1949 Historical Episodes

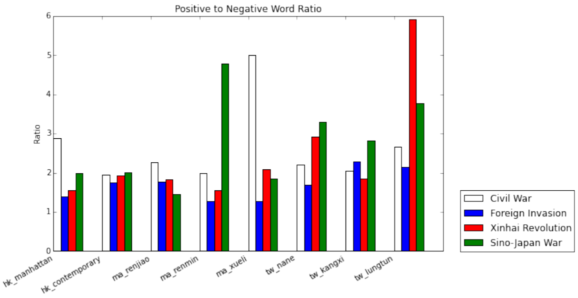


Fig. 9: Positive to Negative Phrase Ratio in 4 major post-1949 Historical Episodes

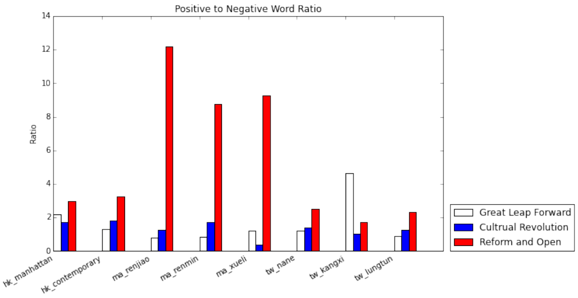


Figure 8 and 9 plot the positive-to-negative phrase ratio of the major historical episodes. Among the pre-1949 events, we do not observe a region-specific pattern but among the post-1949 events, the ratio is noticeably higher in all three Mainland versions of Reform and Open Policy, which is approximately twice the Hong Kong versions and thrice the Taiwanese versions. On the other hand, the ratio is similar across the board for the descriptions on Cultural Revolution and Great Leap Forward.

[12] Lau et al. (2006) found that negative political advertising can stimulate knowledge about the campaign. On the other hand. Herbert et al. (2006) showed that people are more engrossed in processing of the pleasant than of the neutral or unpleasant materials.

[13] The dictionary is available at: http://www.keenage.com/html/c_bulletin_2007.htm.

[14] Table VII and VIII in the appendix also show the top 5 most used positive and negative phrases. 严重 (serious) is the most common negative phrase, and is the top 5 most frequent phrases in all versions. 新 (new) is the most common positive phrase. There is significant overlap among the most frequently used positive and negative phrases in textbooks belonging to the same region. For example, the top 5 most frequently used positive phrases in the Mainland Renjiao and Yuelu versions are identical. In addition, 专制 (dictatorial) and 严重 (serious) are present in all three Mainland versions as top five most frequently used negative phrases. Three positive phrases, and four negative phrases shared the top five between the two Hong Kong versions. In Taiwan, Kangxi and Lungtun share more commonly used phrases, but less so with Nanye.

TABLE III: Number of Positive and Negative Words in Each Version

Region	Version	Positive	Negative	Ratio
Hong Kong	Manhattan	3376	1003	3.4
Hong Kong	Modern	4088	1149	3.6
Mainland	Renjiao	2415	531	4.5
Mainland	Renmin	3721	740	5.0
Mainland	Yuelu	2404	495	4.9
United States	SMC	21171	8367	2.5
Taiwan	Nane	1605	581	2.8
Taiwan	Kangxi	1433	465	3.1
Taiwan	Lungtun	1973	564	3.5
Mainland	old	3217	833	3.9
Taiwan	old	2002	636	3.1

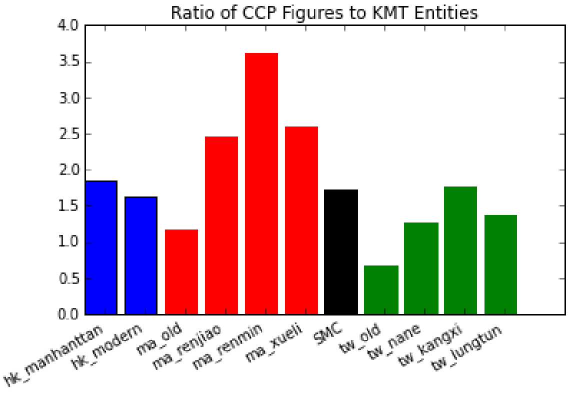
D. *Political Entities*

The occurrence frequency of a political entity can indicate the historical importance of that entity as perceived by the textbook editor. Generally, the more often an entity is mentioned, the more information is provided about that entity. We consider three CCP (Mao, Marx, and CCP) and three KMT entities (Chiang Kai-shek[15], Sun Yat-sen, and KMT) to assess the relative partisan emphasis. Mao Zedong - the central figure of CCP from its foundation until his death - is critical to CCP's authority whereas Chiang Kai-shek - the Director-General of KMT until his death in 1975 - is also a paramount figure. Sun Yat-sen was instrumental in ending the Qing Dynasty and his political philosophy known as the "Three Principles of People," is a cornerstone of the policy adopted by KMT. How often these entities are mentioned could imply support for one partisan perspective over another, as Tumasjan et al. (2010) found that relative number of tweets mentioning a party or candidate is a good predictor of the support that party garners in an election.

Table IV presents the result. There is a large difference in count for Chiang, Marx, and Mao between Mainland and Taiwanese versions. All Mainland versions mention Chiang less than 15 times and Mao more than 45 times, while the opposite holds for Taiwanese versions. Likewise, Marx is mentioned less than 5 times in all Taiwanese textbooks but at least 19 times in the Mainland versions. The other 3 entities receive a more even-handed treatment: the number of mentions of CCP is 113 vs. 103 times in Mainland and Taiwanese versions, whereas Sun is mentioned 30 times on average in both. Interestingly, KMT is mentioned more in Mainland (96 times) than in Taiwan (47 times) versions. The ratio is presented in figure 10. Not surprisingly, Mainland versions have the three highest CCP to KMT ratios. In particular, Renmin mentions CCP and KMT entities in a 3.5 to 1 ratio. Hong Kong and Taiwan textbooks, as well as SMC, have a ratio ranging from 1 to 2, with Taiwanese versions having the lowest ratio among the three regions on average.

There is also a large difference between Mainland and Taiwan pre-reform versions among the CCP entities. CCP, Mao, and Marx are mentioned 152, 62, and 14 times in the old Mainland textbook and 74, 2, and 1 times in the old Taiwan textbook, respectively. However, the difference is less striking for KMT entities. Both Chiang (41 in Mainland, 42 in Taiwan) and Sun (37 in Mainland, 39 in Taiwan) are mentioned about equal number of times. Moreover, KMT is mentioned more in Mainland versions (116 times) than in Taiwan (32 times) in the pre-reform versions. The result signals that CCP entities were more polarizing than the KMT entities before the reforms by term frequencies. After the reform in Taiwan, all CCP entities are mentioned more. On the other hand, all KMT entities are mentioned less after the reform in Mainland.

Fig. 10: Ratio of CCP to KMT Entities Mentions



E. *Word Embedding*

Merely counting the number of times a political entity is mentioned in a version does not reveal how the entity is being described. CCP entities might be mentioned often in Taiwanese textbooks but they might be portrayed unfavorably. To understand how a political entity is being portrayed in more granularity, we study the word embedding. Word embedding is a representation of document vocabulary that aims at quantifying semantic similarities between words based on their distributional properties. The method was originally introduced by Bengio et al. in 2003. The main idea is that words that tend to appear in similar contexts are likely to be related. Mathematically, a word embedding $W : words \to \mathbb{R}^N$ maps words in

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

TABLE IV: Number of Mentions of 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	Yuelu	47	100	19	12	85	21
United States	Search for 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

some text to a high-dimensional vectors. The size of the high-dimensional vectors is predetermined. This method allows us to capture the sentiment surrounding the political entities of interest by computing the distance with their surrounding adjectives in the high-dimensional embedding space trained by the history textbooks from each region.

Implemented with the Python library *Gensim*, we use the Continuous Bag of Words (CBOW) model in Word2vec (Mikolov et al. 2013) to construct word embedding through training of a neural network. The model can be used to predict co-occurrence relationships using the conditional probability of observing the target word given the input context words. 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:

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

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. 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 v_w are two representations of the word w. u_w comes from rows of the input to hidden weight matrix in the neural network, 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 the hidden layer, and the hidden layer to the output matrix. Then to measure the similarity between words, the *Gensim* library computes cosine similarity between a simple mean of the projection weight vectors of the given words and the vectors for each word in the model. We use this method to examine the distance between the political entities studied in the previous section and the adjectives.

Rather than training a model for each publisher version separately, we combine the versions into a corpus by region and train three separate model based on region-specific text to maximize statistical power in explaining cross-region differences in language use. We train each model with an embedding dimension of 500 and a context size of 6. We then search for the closest adjectives around the important political entities in each embedding space.

To facilitate comparison across regions, we calculated the ratio of positive-to-negative phrases using the top 20 surrounding adjectives of the 6 political entities. Table VI shows that Mainland versions have the highest positive-to-negative phrase ratio on all 3 CCP entities. On the other hand, Taiwanese versions have the highest positive-to-negative ratio on Chiang and KMT, while Hong Kong's positive-to-negative ratio tends to fall in the middle. With a value ranging from 0.6 to 1.5, SMC's ratio does not fluctuate a lot across the 6 entities and is generally more balanced. In the appendix, we also presented the cloest adjectives for each of the entity and the word embedding visualizing using the t-SNE technique.

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 adjectives in Hong Kong and Taiwan textbooks. With respect to Chiang Kai-shek, more positive phrases such as 取得胜利 (to get victory) and 固守 (defend tenaciously) are in the list of the closest adjectives in Taiwan textbooks, and noticeably more negative phrase such as 全军覆没 (the whole army is wiped out) in Mainland textbooks.

TABLE V: 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
SMC	1.5	0.6	0.7	1.0	1.0	0.75

V. Conclusion

The growth of nationalism in mainland China, particularly among the young generation, is hard to miss on the internet. Mainland youngsters who lash out online at "disrespectful" public figures and organizations often receive implicit approval from the authority. This presents a striking contrast to the young people in Hong Kong and Taiwan, who are more eager to embrace "Western values" such as democracy and individual freedom. Researchers have suggested that historical education and propaganda apparatus, boosted by the economic success that China enjoyed in the last 20 years, played an important role in mobilizing popular support for the CCP.

This study investigates this premise and provides a quantitative analysis of history textbooks used in these three regions. We showed that mainland history textbooks have a higher adjective and positive-to-negative phrase ratio in post-1949 content, particularly in the description of the Reform and Open policy. Moreover, CCP entities have higher word occurrence frequencies, and are more likely to be described with positive adjectives in mainland textbooks. By analyzing the textbook before and after curriculum reform, we also showed that textbooks used in period associated with more government control have a higher adjective ratio. Post-reform mainland textbooks also have a higher positive-to-negative phrase ratio in post-1949 content. These findings provide more context to the notion that history textbooks in mainland China are often associated with a more subjective narrative that evoke nationalistic sentiments among its readers.

References

[1] S. Adwan and D. Bar-On. *Learning Each Other's Historical Narrative: Israelis and Palestinians*. Peace Research Institute in the Middle East.

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

[3] A. Banfield. *Unspeakable Sentences*. Boston: Routledge and Kegan Paul.

[4] N. Beauchamp. Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 2016.

[5] D. Cantoni, Y. Chen, D. Yang, N. Yuchtman, and Y. J. Zhang. Curriculum and ideology. *Journal of Political Economy*, 125(2):338–392, 2017.

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

[7] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics*, 1:299–305, 2000.

[8] C. Herbert, T. Ethofe, S. Anders, M. Junghofer, D. Wildgruber, W. Grodd, and J. Kissler. Amygdala activation during reading of emotional adjectives — an advantage for pleasant content. *Social Cognitive and Affective Neuroscience*, 4(1):35–49, 2009.

[9] J. Kissler, C. Herbert, P. Peyk, and M. Junghofer. Early cortical responses to emotional words during reading. *Psychological Science*, 2007.

[10] R. Lau, L. Sigelman, and I. Brown Rovner. The effects of negative political campaigns: A meta-analytic reassessment. *The Journal of Politics*, 69(4):1176–1209, 2007.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] B. Qin, D. Stromberg, and Y. Wu. The political economy of social media in china. *Manuscript*, 2017.

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

[19] 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.

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

[21] 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.

[22] J. Savoy. Lexical analysis of us political speeches, 2010.

[23] J. Savoy. Trump's and clinton's style and rhetoric during the 2016 presidential election, 2017.

[24] G. W. Shin and D. C. Schneider. Divided memories - history textbooks and the war in asia, 2011.

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

[26] Y. Theocharis, B. Pablo, Z. Fazekas, S. A. Popa, and O. Parnet. A bad workman blames his tweets the consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication*, 66:1007–1031, 2016.

[27] 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.

[28] 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.

[29] J. Wiebe, R. Bruce, and T. O'Hara. Development and use of a gold standard data set for subjectivity classifications. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, page 246–253, 1999.

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

[31] 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.

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

VI. Online Appendix (Not Intended for Publication)

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	Yuelu	新 (new) 75	科学 (science) 47	和平 (peace) 43	重要 (important) 43	统一 (unified) 34
United States	Search for 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 (enthusiastic) 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 VII: 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 (difficult to) 16
Hong Kong	Modern	严重 (serious) 52	反 (oppose) 40	难以 (difficult to) 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	Yuelu	斗争 (struggle) 27	严重 (serious) 19	专制 (dictatorial) 16	旧 (old) 15	错误 (mistake) 13
United States	Search for Modern China	似乎 (as if) 141	官僚 (bureaucracy) 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 (strate 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 VIII: 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 IX: Mainland 10 closest adjectives

(毛泽东)		KMT (国民党)		CCP(共产党)		Sun(孙中山)		Marx(马克思)	
	Distance	Word	Distance	Word	Distance	Word	Distance	Word	Distance
	0.6510	正面 (positive)	0.6635	重新 (again)	0.5732	腐朽 (rotten)	0.5724	广泛 (extensive)	0.6825
ame)		(positive)		(again)		(rotten)		(extensive)	
正	0.5554	轰轰烈烈	0.5740	正面	0.5494	平均	0.4795	具体	0.6699
of chaos)		(vigorous)		(positive)		(average)		(concrete)	
想	0.5518	大举进攻	0.5341	各阶	0.5469	尊严	0.4627	苦苦	0.6586
ughts)		(launch a large-scale attack)		各阶		(dignity)		(strenuously)	
	0.5435	危急	0.4671	充满	0.5443	另起炉灶	0.4616	大	0.5647
)		(critical)		(be brimming with)		(make a fresh start)	0.4616	(big)	0.5647
	0.4973	不均	0.4570	广泛	0.5207	准确	0.4564	坚定	0.5517
e)		(unequal)		(extensive)		(Accurate)		(firm)	
	0.4922	稳定	0.4506	拨乱反正	0.5142	红 (red)	0.4522	焕然一新	0.5111
t)		(stable)		(bring order out of chaos)	0.5142	(red)	0.4522	(take on an entirely new look)	0.5111
	0.4600	恐慌	0.4490	团结	0.4984	满	0.4484	深远影响	0.5019
t)		(frightened)		(united)		(full)		(have far-reaching influence)	
	0.4565	取得胜利	0.4404	团结一致	0.4801	深入人心	0.4066	良好	0.4785
ve)		(get victory)		(united)		(strike a deep chord in the hearts of the people)		(fine)	
	0.4218	均匀	0.42925	独特	0.4474	旧	0.3868	独特	0.4718
)		(even)		(unique)		(old)		(unique)	
火	0.4160	残暴	0.3887	焕然一新	0.4465	易	0.3815	精神文明	0.4696
rt a prairie fire)		(brutal)		(take on an entirely new look)		(simple)		(spiritual civilization)	

e adjectives in this analysis.

TABLE VI: Most Popular 4-character Adjectives by Publisher Version

Region	Book	Quadgram	Count
Hong Kong	Manhattan	内忧外患 (domestic trouble and foreign invasion)	5
		实事求是 (be true to facts)	4
		前所未有 (unprecedented)	4
Hong Kong	Modern	独立自主 (act independently and of one's own initiative)	14
		内忧外患 (domestic trouble and foreign invasion)	6
		有识之士 (man of insight)	6
Mainland	Renjiao	独立自主 (act independently and of one's own initiative)	8
		百家争鸣 (contention and flourishing of numerous schools of thought)	7
		百花齐放 (flourishing art and literature)	7
Mainland	Renmin	独立自主 (act independently and of one's own initiative)	11
		拨乱反正 (bring order out of chaos)	5
		实事求是 (be true to facts)	11
Mainland	Yuelu	独立自主 (act independently and of one's own initiative)	5
		百花齐放 (flourishing art and literature)	4
		深入人心 (strike a deep chord in the hearts of the people)	3
Taiwan	Nanyi	有识之士 (man of insight)	2
		痛定思痛 (draw a lesson from a bitter experience)	2
		救亡图存 (save the nation from doom and strive for its survival)	2
Taiwan	Kangxi	有识之士 (man of insight)	3
		百家争鸣 (contention and flourishing of numerous schools of thought)	2
		苛捐杂税 (exorbitant and multifarious taxes and levies)	2
Taiwan	Lungtun	贪污腐化 (corruption and degeneration)	5
		国计民生 (national economy and people's livelihood)	2
		救亡图存 (save the nation from doom and strive for its survival)	2

TABLE XI: Taiwan 10 closest adjectives

Chiang (蔣介石)		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)		(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)	

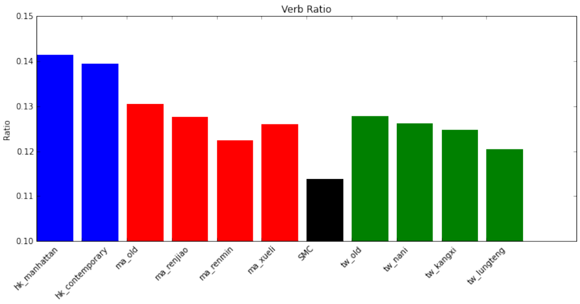
Notes: We use Jiena's part-of-speech tagging to define adjectives in this analysis.

TABLE XII: SMC 10 closest adjectives

Chiang (蔣介石)		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	
两败俱伤	0.7250	犀利	0.7452	桀傲不驯	0.8895	仅仅只是	0.8367	悻悻	0.8157	宽大为怀	0.8433
(neither side gains)		(sharp)		(arrogant and unyielding)		(only)		(angry)		(be magnanimous)	
日经	0.6794	才 0.7360	乱事	0.8884	胆怯	0.8317	悻悻然	0.7838	适宜	0.8295	
(already)		(actually)		(troubled event)		(cowardly)		(angry)		(appropriate)	
骨肉相残	0.6391	独秀	0.7073	争执不下	0.8873	为所欲为	0.8165	偏僻	0.7781	豪言壮语	0.7558
(fratricidal fighting)		(duk-sau) (Name)		(neither could convince the other)		(do as one pleases)		(far away)		(heroic utterance)	
不惜一切	0.5726	未公开	0.7022	亲自	0.8780	尖	0.8160	娇弱	0.7380	马首是瞻	0.7384
(at all cost)		(undisclosed)		(personally)		(sharp)		(weak)		(follow somebody's lead)	
宽宏	0.5724	从先	0.6996	少壮	0.8744	最早	0.8129	悻然	0.7280	宽大	0.7180
(magnanimous)		(in the past)		(young and strong)		(earliest)		(angry)		(spacious)	
左支右绌	0.5179	深谋远虑	0.6940	亲密	0.8724	一己之私	0.8098	活泼	0.7130	冰消瓦解	0.6898
(be in straitened circumstance)		(think deeply and plan carefully)		(close)		(pursue one's own ends)		(live)		(disintegrate)	
争执不下	0.5079	经直	0.6782	暂	0.8719	纯美	0.7736	险些	0.7077	正巧	0.6793
(neither could convince the other)		(straight)		(temporarily)		(pure beauty)		(almost)		(chance to)	
反共	0.4962	整肃	0.6771	溃不成军	0.8707	浴火重生	0.7609	最强	0.6976	欢愉	0.6598
(anti-communism)		(enforce)		(be defeated and flee in great disorder)		(rebirth)		(strongest)		(happy)	
正前	0.4917	心知肚明	0.6689	毋须	0.8695	正是	0.7570	无足轻重	0.6969	浅薄	0.6339
(in front)		(be well aware)		(unnecessary)		(as is)		(be of little significance)		(shallow)	
桀傲不驯	0.4607	拙劣	0.6667	逐级	0.8685	浓烈	0.7550	大不相同	0.6847	耀眼	0.6211
(arrogant and unyielding)		(clumsy and inferior)		(step by step)		(strong (flavor))		(very different)		(dazzle)	

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

Fig. 11: Verb Ratio

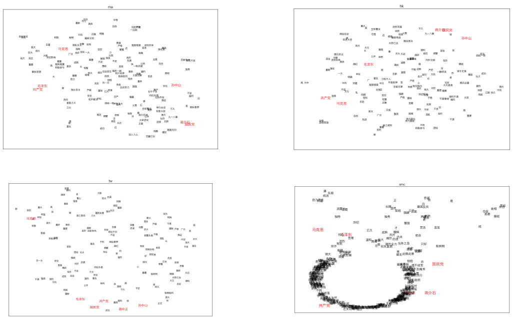


A. *Word Embedding Visualization*

We can visualize each embedding space trained by the model 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[16]. By visualizing them in a 2-dimensional plane, we can illustrate the semantic distance between the political entities themselves.

Figure 12 shows the visualization of our embedding space. While the relationship between Marx, CCP and Mao is similar to that of Sun and Chiang and KMT in Mainland textbooks, the two clusters are far from each other. In Hong Kong, CCP and Marx are close to each other while Mao appears to be further apart. Mao is closer to the KMT, Chiang and Sun cluster. Finally in Taiwan, Mao, CPP, KMT, Chiang and Sun are clustered and Marx alone is far away. This echoes with the previous finding that Marx is scarcely mentioned in Taiwan textbooks.

Fig. 12: Word embedding visualization of Mainland 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)

[16] 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 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.