

Negotiating a Century of Humiliation

How Baidu Baike and Chinese Language Wikipedia Differ in Writing Modern Chinese History (1840-1949)

Jingwen Liu

June 12, 2020

1 Introduction

How does censorship affect the collective writing of history? Critical historical moments often serve as pivots that frame and shape people's understanding of the present. With crowdsourcing online encyclopedias increasingly become publicly available information and knowledge base, Internet users have the opportunity to write and access history besides textbooks and academic works. However, such practices are constrained in China under censorship regulations and strong control over historical narratives by the Chinese Communist Party (CCP).

In this paper, I study how key events and entities in Modern Chinese history are written differently on Baidu Baike and Chinese language Wikipedia, the top 2 online encyclopedia projects in Chinese language, by analyzing the text corpora of entries¹. With different levels of exposure to censorship, I argue that the information on Baidu Baike is a result of an unequal negotiation between users and the authority as opposed to an equal negotiation among users on Chinese language Wikipedia (Chinese Wikipedia hereafter). I find that Baidu Baike is more subjective than Chinese Wikipedia in depicting events during the Century of Humiliation, and Marxist Historiography and the CCP narrative are more prevalent in Baidu Baike's accounts. Moreover, the disagreement between Baidu

¹A text corpus refers to a collection of structured text. The corpora in this case are entry pages on Chinese language Wikipedia and Baidu Baike. An entry refers to an online encyclopedia page dedicated to one entity.

Baike and Chinese Wikipedia is driven by whether the CCP was present in an event, while the agreement between them is determined by traumaticity in an event.

Scholars of censorship have paid attention to evidence from Wikipedia in how the block by the Great Firewall of China affects information consumption and contribution, i.e. readership and editorial contributions. The intermittent blocks caused a large decrease in contribution levels of all Wikipedia users including those in Hong Kong and Taiwan who are not affected by the Firewall (Zhang and Zhu, 2011). The complete block in 2015 also led to a significant negative impact on page views of Chinese Wikipedia (Pan and Roberts, 2019). Moreover, Hobbs and Roberts (2018) shows that there is a large increase of views in political and historical entries on Wikipedia when there is a spike of people who jump the Firewall. However, Baidu Baike, Wikipedia's alternative inside the Great Firewall, has received little attention. Ng (2013) first studied the coverage difference of Baidu Baike and Chinese Wikipedia by automatically detecting a list of mutually exclusive entries that only appear on one platform and found that most entries omitted on Baidu Baike contain either political sensitive information or adult contents. Few studies have further looked at substantive differences in the text corpora and information quality on Baidu Baike and Chinese Wikipedia, let alone historical knowledge in specific.

Historical knowledge is interesting to study not only because it shapes understandings of the present with political implications, but also because Baidu Baike and Chinese Wikipedia as online collaborative encyclopedias feature a space for people to negotiate and develop collective memories that form the basis of historical narrative. Nevertheless, existing studies on historical education and narrative manipulation have been focusing on the development of official discourses (Wang, 2014; Cohen, 2002; Kaufman, 2010; Coble, 2007) and textbooks (Cantoni et al., 2017; Shin and Sneider, 2011). Lam and Lin (2017) analyzed high school history textbooks in mainland China, Hong Kong and Taiwan using quantitative text analysis, and found mainland China textbooks have more adjectives overall and a higher positive-to-negative phrase ratio in post-1949 content. Studying the writing of modern Chinese history on Baidu Baike and Chinese Wikipedia adds to the source of evidence on the differentiating depictions of Modern Chinese history that serve

political ends.

This study attempts a quantitative examination of the entry texts on the Century of Humiliation on Baidu Baike and Chinese Wikipedia. The differences arising from the comparison can confirm how a “censored” narrative as is represented by Baidu Baike diverges from a general “region-less” narrative negotiated by decentralized Chinese-speaking users as is represented by Chinese Wikipedia. The similarities can shed light on where the consensus among all Chinese-speaking users stands.

This paper proceeds as follows. In the next section, I propose a framework of information negotiation on Baidu Baike and Chinese Wikipedia due to different levels of censorship exposure and discuss the expected differences in the context of the Century of Humiliation. Section 3 describes the text corpora dataset I use and the variables acquired and constructed in the study. Section 4 presents the empirical findings, which is followed by a discussion and conclusion.

2 Background and Theoretical Framework

Collective memory is constructed past through complex communication among a social group shaped by the concerns of the present (Halbwachs et al., 1992; Assmann and Czaplicka, 1995), which forms the mental representation of historical events. Before the World Wide Web, the imagined communities are facilitated by printed media and broadcasting nested in languages, without members meeting each other (Anderson, 2006). The collaborative modern Internet environment and wikis provide a tangible and traceable space for the formation of collective memory through interactions. Wiki projects, as sets of web-based interlinking pages that construct knowledge bases, feature instant, observable and collaborative creating and editing by all users. Wikipedia therefore becomes a memory place where discussions and negotiations over memorable elements fabricate collective memory (Pentzold, 2009). Baidu Baike also engages in the same process of collective memory construction. This makes the negotiated information and knowledge displayed on Baidu Baike and Chinese Wikipedia the perfect source to study how history

is framed and remembered.

In this section, I first describe the highly restrictive internal mechanism on Baidu Baike in compliance with censorship regulations and how it contrasts Chinese Wikipedia and yields different information negotiation patterns. Then, I discuss the importance of the Century of Humiliation to China and the CCP, and propose the expected differences in the corpora.

2.1 Information Display and Control on Baidu Baike and Chinese Wikipedia

Table 1: Basic Information

	Chinese Wikipedia	Baidu Baike
Type	Non-profit	Commercial
Year of Establishment	2001	2006
Total Entries (millions)	1.1	16.7
Total Edits (millions)	58.2	158.3
Total Users (millions)	2.9	7.0

Source: publicly available information on baike.baidu.com and zh.wikipedia.org/wiki/Wikipedia:%E7%BB%9F%E8%AE%A1, as of April 2020.

Wikipedia is a collection of wikis in different languages ran by a non-profit organization Wikimedia, and Chinese Wikipedia is one wiki project that invites all Chinese-speaking users to collaboratively contribute to the knowledge base. Baidu, a Chinese Internet giant known for its search engine, established an alternative collaborative online encyclopedia Baidu Baike² in mainland China five years later than Chinese language Wikipedia in 2006. Today, the fast-growing commercial Baidu Baike has become 15 times larger in volume than Chinese Wikipedia. While media marketization in China lifts the information credibility to ordinary citizens (Stockmann, 2012), all Internet Content Producers are subject to censorship regulations in that they are required to remove content at the

²Baike is the mandarin Chinese word for “encyclopedia”.

direction of the government(Miller, 2018). Different internal mechanisms and external interventions therefore determine that there will be difference in information generated and displayed on Baidu Baike and Chinese Wikipedia. *If Wikipedia represents the negotiation results among decentralized contributors themselves of divergent backgrounds, the negotiation behind Baidu Baike is between individual contributors and centralized content control and censorship enforced by Baidu.*

Table 2: Comparison of Editorial Policies

	Chinese Wikipedia	Baidu Baike
Display Policy	Immediately effective without any screening; ex post patrol by Elite Users	ALL edits go through ex ante oversight by Baidu
Protection Policy	Edit-warring	Edit-warring, Politically sensitive, Scientific
Editing Threshold	None, IP address on record	Real-name registration required
Material Incentives for Editors	No	Yes
Elite Team	Voted by users	Appointed by Baidu
Number of Elite Users	79	412

Display Policy As a loyal practitioner of the wiki principle of open and traceble collaboration, Chinese Wikipedia can be edited by anyone at any time, and the edits will take effect and be immediately displayed at the entry page. On the contrary, on Baidu Baike, ALL new edits are required to go through a review process maintained by both Baidu administrators and bots, which can take as long as 16 days to be approved and displayed. After making proposed edits in the editing interface of Baidu Baike, users have to submit the request along with a short reason for editing that assists the review.³ Users will be notified of the review results on their account dashboards or via text message or WeChat if opted in. The often inefficient review process receives wide complaints on Baidu Forum,

³There are pre-filled options such as error correction, content enlarge, internal links modification, redundancy removal, etc. Users can also choose to write their own explanations.

and is seen by both Baidu users and Wikipedia users as a black box⁴. This compulsory review process grants Baidu a comprehensive *ex ante* control over what to display on Baidu Baike entry pages.

Wikipedia, because of its openness, acknowledges that it “is not a reliable source” itself. Wikipedia therefore is dedicated to promoting and practicing three core content values: neutral point of view, verifiability, and no original research. Practicing neutral point of view, Chinese Wikipedia strives to avoid regional-centric (避免区域中心)⁵ expressions, e.g. use Taiwan or the Republic of China (ROC) not Taiwan province, and use the Gregorian calendar not ROC calendar. Baidu Baike is much less sophisticated in making editing rules, and the emphasis has been that Baidu Baike is free of commercials.

Protection Policy To better ensure information quality, Chinese Wikipedia implements protection policies when there are vandalism, disruptive editing, or damaging disputes. This is a natural development in the open environment where edits are instantly effective and disputes can easily be toxic. After a volunteer administrator locks down an entry that is disrupted by back-and-forth edits, all discussions are directed to the WikiTalk page associated with the entry where parties can try to agree before the entry reopens.

Nevertheless, Baidu Baike engages with more restrictive measures under the cover of protection policies. In addition to protecting entries involved with “edit warring” like Wikipedia, Baidu Baike also locks down politically sensitive entries to “remain the completeness of the established content”⁶. and scientific entries to “remain the precision and completeness of the content provided by experts”.⁷ This is a powerful pre-cautious control over the information displayed on Baidu Baike.

⁴Zhang, Jane. “How Baidu built an encyclopedia with 16 times more Chinese entries than Wikipedia,” SCMP. Nov 20, 2019. Accessed at <https://www.scmp.com/tech/big-tech/article/3038402/how-baidu-baike-has-faced-against-wikipedia-build-worlds-largest>.

⁵See <https://zh.wikipedia.org/wiki/Wikipedia:%E9%81%BF%E5%85%8D%E5%9C%B0%E5%9F%9F%E4%B8%AD%E5%BF%83>

⁶Baidu Baike has made clear of its policy that all edits for entries on political leaders and government agencies are subject to additional reviews. See https://baike.baidu.com/reference/8646268/eab4j6A2-N0xZDBzieRUwX0j7BKd48nsXh4Zzx_uXBn-EM8TAzbJk94Rm88qLUPNhyetbo1Rbual1v2TGpYoHEc.

⁷A scandal that might lead Baidu to be extra-cautious on scientific entries is the incident of Wei Zexi. Wei Zexi, a college student who died of a rare cancer, accused Baidu of promoting falsified information on less proven treatments. See more at <https://www.nytimes.com/2016/05/04/world/asia/china-baidu-investigation-student-cancer.html>

Editors Consistent to the instant display, there are no threshold for editing practices on Chinese Wikipedia. No registration is required, and the IP addresses of anonymous users are recorded in editing history. Whereas all editing activities on Baidu Baike now require an account with real-name registration, which is an implicit signal to users to make higher quality and “correct” edits. Many entries only open their editing access to high-level Baidu accounts that have high editorial approval rates. Baidu Baike is also able to incentivize users to enlarge the scope and enrich the coverage of knowledge with material rewards. Products ranging from cooking oil to iPhones be redeemed with wealth points gained from editing activities. For example, one approved edit accounts for one wealth point, and six wealth points approximately worth one RMB.

Elite Team Both platforms have elite teams whose member possess higher level authorization to manage entries. Chinese Wikipedia administrators, who are elected by fellow Wikipedians, voluntarily patrol the content, reverse poor edits, and are the only ones authorized to manage protected pages. As of May 2020, there are 79 administrators and 174 new page reviewers on Chinese Wikipedia. Baidu Baike Elite Team (Kedoutuan, 蝌蚪团) is “directly led and managed” by Baidu. Baidu employees, who can be identified by their online ID, guide elite members in creating new entries and distribute souvenirs such as T-shirts and mugs with the Elite Team symbol to elite members. As of May 2020, there are 412 members in Baidu Baike Elite Team ⁸, who are granted authorizations such as combining synonymic entries and overwriting poor-quality edits. All of their activities, however, still subject to official reviews by Baidu before taking effect.

In sum, Chinese Wikipedia is operated by established content policies and ex post decentralized patrol. Engaging ex ante oversight, Baidu asserts overall control of which content initiated by users to be displayed on Baidu Baike. These internal mechanisms indicate how Baidu Baike is subject to censorship that is achieved by an unequal information negotiation on an online encyclopedia.

⁸Elite Team members receive a monthly “salary” of 6000 wealth points, which equal about 100RMB according to rewards exchange standards in Baike Mall.

The Great Firewall also contributed to creating a dichotomized Chinese language online encyclopedia space in terms of both editing practices and readership as an external intervention factor. Access to Chinese Wikipedia from mainland China had been intermittently blocked by the Great Fire Wall from 2004 to 2008 for a total of at least 1,023 days and selectively blocked by sensitive pages from 2008 to 2015, until it was completely blocked in September 2015.⁹ Liao (2014) documented that Baidu Baike enjoyed a rapid growth of users from 2006 to 2008 that formed its basis, when Chinese Wikipedia was constantly blocked concurring with the rapid growth period of general Internet user in China. Moreover, before the block, Chinese Wikipedia is a collective space for all Chinese speakers in Mainland China, Taiwan, Hongkong, Macao, Malaysia, Singapore, and other places. From the perspective of collective memory construction, Chinese Wikipedia becomes a missed opportunity for mainland China users to engage in the communicative phase of forming a collective memory in the Chinese-speaking community, which is a side-effect of the Firewall's efforts in blocking uncensored information consumption.

Chinese Wikipedia yields communicative and negotiated results among equal and decentralized contributors. As a result of poor content policies and compulsory ex ante oversight, Baidu Baike is driven by an authority in an unbalance negotiation between Baidu and contributors, especially for sensitive entries. In this sub-section, I have laid out how internal mechanism and external interventions can influence the knowledge generated and displayed on Baidu Baike and Chinese Wikipedia. Next, I discuss the expected differences in the context of the Century of Humiliation.

⁹In May of 2015, Wikipedia required that users use the encrypted versions of its pages, making it too difficult for the Chinese censorship apparatus to selectively block particular pages. Therefore all of Chinese language Wikipedia has been blocked since May 19, 2015. See Welinder, Yana, Victoria Baranetsky, and Brandon Black. "Securing Access to Wikimedia Sites with HTTPS," *Wikimedia Blog*, June 12, 2015, <https://blog.wikimedia.org/2015/06/12/securing-wikimedia-sites-with-https/>, and Oberhaus, Daniel, "Wikipedia's Switch to HTTPS Has Successfully Fought Government Censorship," *Motherboard*, May 26, 2017, <https://bit.ly/2T5aEWm>.

2.2 Official Memory, Historical Narrative, and Collective Memory on Online Encyclopedia

In the context of collective memory, the identity of large groups are in particular established and bolstered through “chosen trauma” and “chosen glory” (Volkan, 1997). According to Volkan, traumas introduce a transgenerational sense of powerlessness and humiliation that put a group under a shared hat of victimhood, while glories introduce feelings of success and triumph. The reactivation of both can fuel the rise of self-esteem.

The Century of Humiliation incorporates both chosen traumas and chosen glories relevant to China’s modernization and establishment of national identities. It refers to the historical period from the First Opium War in 1840 to the establishment of the People’s Republic of China in 1949. Facing continuous defeats that brought territory and sovereignty loss, the traditional China-centric views were devastated and the Qing empire came to an end. During the turbulent but somewhat politically vibrant period, national humiliation as a phenomenon was constructed and promoted by various political forces including the Kuomintang (KMT) and the Chinese Communist Party (CCP) in efforts to mobilize Chinese nationals against foreign invaders.

Unlike liberal democracies, authoritarian regimes impose strong control over the official memory, and strictly distant private memory from public memory if the former is allowed at all (Jović, 2004). The CCP, who has always manipulated propaganda and history education in favor of its legitimacy and power withholding, is no exclusion,. The Party has put great emphasis on telling the right story of its early history and how it came to power. In the leaked Central No.9 Document, a briefing and guideline on ideological control, the Party regard historical nihilism as a key field for ideological struggle. Historical nihilism, according to the CCP, denies that the CCP has been the only successful leader of Chinese revolution and socialism is deterministic for China, and hurts the legitimacy of the Party.

Because of the CCP’s incentive to control the historical narrative in combination with the editorial mechanisms of Baidu Baike, I expect Baidu Baike entries to be more subjective than Chinese Wikipedia entries. As Baidu Baike employs the ex ante review, the centralized mechanism delays and discourages interactive conversation among users that

could adjust for subjectiveness. The reviewer would also be less aware of the context. Negotiation among users on Chinese Wikipedia, guided by impartial values, promotes consensus building with less subjectivity.

Hypothesis 1 *Baidu Baike entries show more subjectivity than Chinese Wikipedia entries on the Century of Humiliation.*

Hypothesis 1a *Entries that have more unique editors are less subjective.*

Two additional sources of constraints allow for more variations in discrepancies of entry contents on the Century of Humiliation on Chinese Wikipedia and Baidu Baike, given different oversight and negotiation mechanisms and the Great Firewall.

To begin with, the CCP is constrained by its own legitimacy concerns and policy agendas. In the Mao era, the CCP was framed as the only glorified leadership that saved China from the corrupt Qing feudal government and KMT and foreign invaders, and gained national independence in the anti-feudalism and anti-imperialism struggles (Wang, 2014). This leaves a legacy of an analytical framework by Marxist historiography that frequently uses a class struggle narrative. After the 1989 Tiananmen Protests, the CCP launched a patriotic education campaign aiming at establishing strong national identities, which shifted the emphasis of the narrative from CCP-centric to Chinese nation's struggles with outside forces, i.e. from a victor to a victim (Wang, 2014). Under the new narrative, the KMT is given proper credits in the Anti-Japanese War, whereas the CCP claimed credits for all critical victories before.

Hypothesis 1b *Baidu Baike entries reflect Marxist historiography.*

Hypothesis 2 *Baidu Baike favors the CCP with positive sentiment, while resonates with Chinese Wikipedia in the period of Anti-Japanese War.*

Second, history during the Century of Humiliation is too recent and salient to be completely dominated by one official discourse. It is a shared memory of the Chinese speaking community where Taiwan and Hongkong are stakeholders in many of the major events, which has enabled dynamic contributions to Chinese Wikipedia. Traumas are

intergenerational because the losses need to be externalized to be resolved (Volkan, 1997). Therefore, it is also impossible to eliminate private memory from public memory on the Century of Humiliation in mainland China as the first-hand witnesses and their offspring still have the voice.

Hypothesis 3 *Baidu Baike and Chinese Wikipedia are more similar on traumatic events.*

3 Data

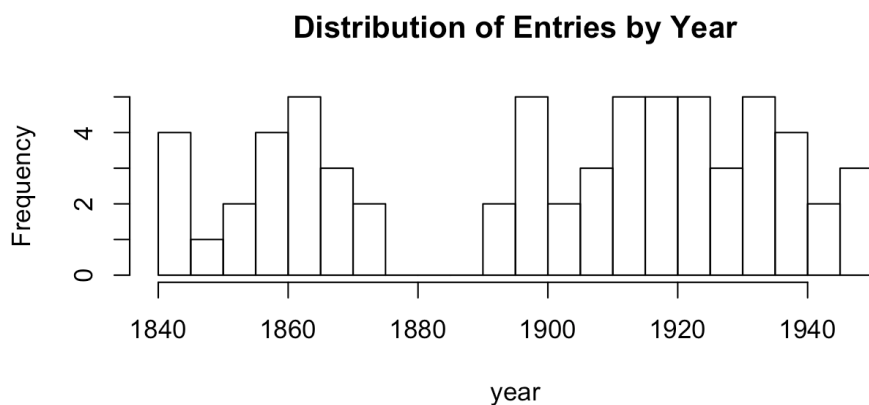


Figure 1: Histogram of Entries by Year

I scraped 132 paired entries from Chinese Wikipedia and Baidu Baike as of March 2020, which contain major historical events and entities in China from the period of 1840 to 1949. The list of events and entities is compiled based on the timeline of major historical events in the appendix of Renjiao (人教版) mainland China high school history textbook, which is manually supplemented based on *The Rise of Modern China* by Immanuel C. Y. Hsu to avoid obvious negligence (See Appendix for a complete list of entries). I constrain the the scope of the data in this study to 1) major historical events and entities which are 2) covered by both Baidu Baike and Chinese Wikipedia for three reasons. First, only major historical events and entities that recieve certain level of attention can draw enough contributions on Baidu Baike and Chinese Wikipedia. Second, since my focus is on the differences in content of entries, I make sure that the entires are comparable by matching events and entities. Therefore, I will not discuss the coverage differences of Chinese

Wikipedia and Baidu Baike.¹⁰ Third, if I include entries by the categorization system or internal links embedded on Chinese Wikipedia and Baidu Baike, it would be hard to determine the boundary and eliminate irrelevant entries and noises in an automatic process.

Figure 1 shows the distribution of entries by event year.¹¹ Notably, there is an “empty window” in this dataset between 1875 and 1890. In specific historical circumstances, oblivions spring narratives in national biographies (Anderson, 2006). Recovering the context, this relatively “eventless” period is during the later phase of the self-strengthening movement (自强运动) by the Qing government, when the modernization process continued in infrastructure, heavy industry, and military. However, as there were no major new defeats and concessions to foreign powers, the period becomes partly neglected in the narrative of the Century of Humiliation.

Table 3: Examples of Entry Name Differences

Chinese Wikipedia	Baidu Baike
the 715 Incident (七一五事变)	the 715 Counter-Revolutionary Coup (七一五反革命政变)
Battle of Pingxing Pass (平型关战役)	Great Victory of Pingxing Pass (平型关大捷)
Battle of Taierzhuang (台儿庄战役)	Great Victory of Taierzhuang (台儿庄大捷)

While the paired entries are on the same matters, it is worth pointing out that some of the entry names are different on Chinese Wikipedia and Baidu Baike. Those differences already reflect Chinese Wikipedia’s neutral standpoint and Baidu Baike’s CCP perspective. For example, Wikipedia lists the 715 Incident, which refers to the purge of communists by the KMT Wuhan government led by Wang Jingwei on 15 July, 1927, as the 715 Incident (七一五事变) and clarifies in the first sentence that it is referred

¹⁰There are indeed coverage differences which are relevant to both censorship and collective forgetting. One example relevant to the historical period in this study is that the Dungan Revolt (1862–1877, 同治陕甘回变) is covered by Chinese Wikipedia but not Baidu Baike.

¹¹The year of entries of important figures are determined by their most prevalent activities. For example, the year for Lin Zexu (林则徐, known for his role in the First Opium War) is coded as 1840. For events that lasted for more than a year, the middle year is taken as the event year.

by the KMT as Wuhan-Communist Split (武汉分共) and the CCP as the 715 Counter-Revolutionary Coup (七一五反革命政变). Baidu Baike takes the later name the 715 Counter-Revolutionary Coup (七一五反革命政变) only. Moreover, names on Baidu Baike tell the framing by the CCP. Battle of Pingxing Pass (平型关战役) is a part of the Battle of Taiyuan during the Anti-Japanese War, which was a defeat that contributed to the Japanese’s access to coal in Northwestern China. However, the Eighth Route Army of the CCP received one of its first victories in an ambush attack that slowed down Japanese troops during the Battle, which becomes the focus of Baidu Baike as Great Victory of Pingxing Pass (平型关大捷). Battle of Taierzhuang (台儿庄战役), the first major victory of the National Army (国军) against the Empire of Japan, is also framed on Baidu Baike as Great Victory of Taierzhuang (台儿庄大捷) emphasizing the collective efforts of resistance against Japanese invasion.

To preprocess the corpora in preparation for quantitative analysis, I first converted all texts into simplified Chinese ¹². Unlike English, there are no natural spaces between words in Chinese. In order to separate the text into meaningful phrases, I segmented the text into word tokens with Jieba, a Python Chinese word segmentation module. In addition to the text corpora, I acquired metadata around each entry including numbers of edits, unique editors, and views and coded each event according to its historical period and features. Table 4 shows summary statistics for all variables.

3.1 Dependent Variables

Adjectives The number of adjectives proportional to the total entry length is a main measurement for subjectiveness. I use a part-of-speech tagger in Jieba to detect adjectives. The tagger works with a build-in dictionary that records the category of words. I supplemented the default dictionary with a list of words specific to Chinese modern history to improve its accuracy.

¹²Certain Chinese Wikipedia entries are partially in traditional Chinese.

Table 4: Summary Statistics

Variable	N	Mean	St. Dev.	Min	Max
<i>Corpora</i>					
Length(Tokens)	264	5,394.22	4,689.80	132	27,433
Number of Adjectives	264	91.99	84.02	0	403
Number of Nouns	264	1,629.20	1,400.48	44	7,889
Proportion of Adjectives	264	0.02	0.01	0.00	0.06
Proportion of Nouns	264	0.31	0.03	0.23	0.43
Sentiment Score	264	0.003	0.01	−0.01	0.02
Cosine Similarity	132	0.26	0.17	0.02	0.80
<i>Entry-level Characteristics</i>					
Wikipedia	264	0.50	0.50	0	1
Number of Edits	264	358.9	570.73	5	5,839
Number of Editors	264	157.43	201.14	2	1909
Number of Views	264	2,461,300.00	5,070,007.00	108	41,977,740
Republican Period	264	0.36	0.48	0	1
Anti-Japan Period	264	0.08	0.28	0	1
CCP-related	264	0.20	0.40	0	1
<i>Editor Characteristics</i>					
Lifetime Number of Edits	15,684	3323.6	29583.21	1	2,771,510
Number of Entries Edited in Corpora	24,259	1.61	3.48	1	126
<i>Baidu Baike Editor only</i>					
Edit Approval Rate	11,353	0.79	0.21	0	1
Wealth Points	11,353	26,563.73	326,356.30	0	14,158,780
<i>Chinese Wikipedia Editor only</i>					
Elite Status	12,777	0.01	0.11	0	1
Anonymous	12,777	0.66	0.47	0	1
Blocked	12,777	0.02	0.14	0	1

Sentiment Employing dictionary-based method, I use the Augmented NTU Sentiment Dictionary (Wang and Ku, 2016) to calculate sentiment scores for each entry. After matching the words in the segmented corpora with the ANTUSD, a sentiment index is generated by taking a weighted average of individual sentiment scores by the length of tokens. The sentiment index constructed in the ANTUSD ranges from -1 to 1 , where 0 represents neutrality. In addition, I count the number of positive words and negative words respectively matched in the dictionary and calculate a positive to negative word ratio as an indicator for the range of sentiment.¹³

Similarity I measure the similarity of paired Baidu Baike and Chinese Wikipedia entries by cosine similarity. It is calculated based on the document-feature matrix where each row is an entry and each column is a unique word in the corpora. Vectors of feature frequencies for paired entries are then compared for similarity. This measurement takes into account all words in a given entry.

3.2 Independent Variables

Features of the entries In addition to a key dummy indicating whether it belongs to Chinese Wikipedia and Baidu Baike, each entry comes with the number of previous edits and the number of views it has received. On average, Baidu Baike entries are longer and viewed by more people but have less previous edits than Chinese Wikipedia.

Features of the events The events targeted by the entries are coded in three dimensions: time period, key actor, and emotion. The Century of Humiliation in China can be divided into four periods: 1) Late Qing (1840-1911); 2) Republican (1912-1936); 3) Anti-Japanese War (1937-1944); 4) Civil War (1945-1949). For key actors, events with the presence of the CCP are identified. In addition, events that involve territory and sovereignty loss, defeats, and casualties are coded as traumatic events.

¹³See Appendix for a list of the most positive and negative words appeared in this corpora.

4 Empirical Analyses

4.1 Subjectivity and the CCP Narrative

I begin by testing Hypothesis 1 that expects Baidu Baike to be more subjective than Chinese Wikipedia in writing the Century of Humiliation. For one thing, Baidu Baike will engage in more subjective assertions and absorbs the official CCP historiography (Hypothesis 1b). For the other, Baidu’s information control mechanisms weaken the neutralizing effect of negotiation among users, and therefore adds subjectivity to the information displayed (Hypothesis 1a). As is discussed in Section 3, the degree of subjectivity is measured by the number of adjectives in the corpora. The level of negotiation in this case is proxied by the editor to edit ratio for each entry. Intuitively, a higher editor to edit ratio indicates more unique contributions to an entry, which means the information is less dominated and therefore less subjective.

I use a negative binomial model since the dependent variable is the count of adjectives. Table 5 shows the results. The baseline model in Column (1) confirms that there are more adjectives in Baidu Baike entries than Chinese Wikipedia, holding constant the total number of tokens (entry length). Column (2) adds in controls for number of previous edits and the editor to edit ratio. While the number of adjectives on Baidu Baike is still larger, the negative coefficient for the editor to edit ratio confirms that more unique contributions leads to less subjectivity. Figure 2 shows the expected number of adjectives on Baidu Baike and Wikipedia from 1000 simulations based on the model in Column (2). Holding entry length, number of edits, and editor to edit ratio at their medians, Baidu Baike would use 35 more adjectives than Wikipedia, which is equal to half a standard deviation of the Wikipedia adjective number mean.

In Column (3) of Table 5, I include the dummies of event features in the model, including historical periods, CCP presence, and traumaticity. While the number of adjectives used do not significantly differ across periods, entries with the presence of the CCP as a key actor have more adjectives. However, when an interaction term with Baidu Baike is added in Column (4), the results show no differential marginal effects of CCP events on

Table 5: Negative Binomial Models for Adjective Counts

	<i>Dependent variable:</i>				
	count of adjectives				
	(1)	(2)	(3)	(4)	(5)
Baidu Baike	0.561*** (0.064)	0.718*** (0.088)	0.728*** (0.087)	0.756*** (0.094)	0.233 (0.263)
Length	0.0002*** (0.00001)	0.0002*** (0.00001)	0.0002*** (0.00001)	0.0002*** (0.00001)	0.0002*** (0.00001)
Edit		0.00003 (0.0001)	0.00004 (0.0001)	0.00004 (0.0001)	−0.00003 (0.0001)
Editor/Edit		−0.602*** (0.221)	−0.615*** (0.221)	−0.592*** (0.221)	−1.172*** (0.360)
Baidu*Editor/Edit					0.870* (0.446)
Republican			−0.051 (0.074)		
Anti-Japan			−0.140 (0.133)		
Civil War			−0.270 (0.175)		
CCP			0.187** (0.087)	0.174* (0.102)	
Baidu*CCP				−0.112 (0.143)	
Truamatic			−0.037 (0.097)		
Constant	2.923*** (0.057)	3.227*** (0.130)	3.223*** (0.138)	3.175*** (0.134)	3.524*** (0.195)
Observations	264	264	264	264	264
Log Likelihood	−1,288.106	−1,284.418	−1,281.669	−1,282.744	−1,282.818

Note:

*p<0.1; **p<0.05; ***p<0.01

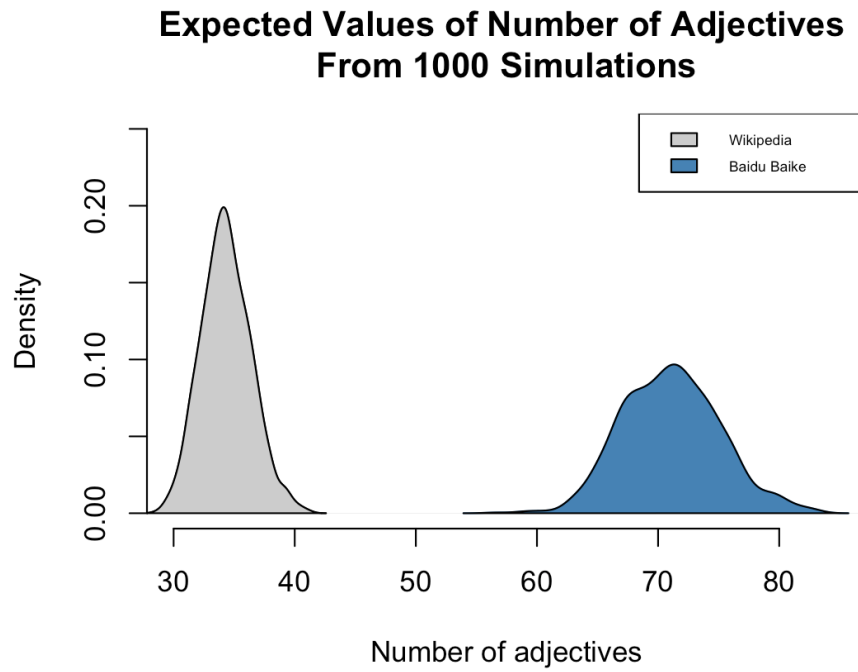


Figure 2: Simulation of Expected Adjective Counts by Type

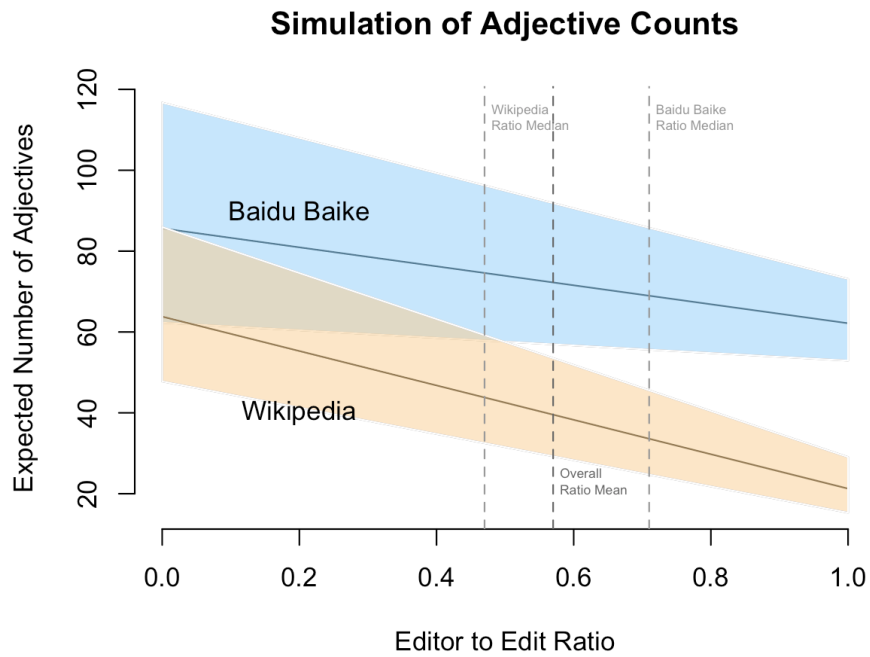


Figure 3: Simulation of Expected Adjective Counts, Interaction with Editor/Edit

Baidu Baike and Chinese Wikipedia.

Finally, I look at the interaction of editor to edit ratio with the platform indicator Baidu Baike. Figure 3 shows the predictions of adjective numbers on Baidu Baike and Wikipedia interacted with the editor to edit ratio. We have learned that the editor to edit ratio is negatively correlated with the number of adjectives, as more individual contributions tend to neutralize the common knowledge. Nevertheless, the additional effect of individual contributions on objectivity is weaker on Baidu Baike than on Wikipedia. In other words, more unique editors on Baidu Baike would contribute less to the overall objectivity of entries than on Chinese Wikipedia. This result implicitly suggests that the underlying censorship mechanism might be damaging the information generation process on Baidu Baike, in that it prevents direct interaction and negotiation among users.

Now that we know Baidu Baike uses more adjective than Chinese Wikipedia in depicting the historical events and entities, in order to make sense of what the adjectives are and what kind of subjective assertions they imply, I run the Structural Topic Model (STM) on the corpora of adjectives, i.e. only adjectives in the segmented text included in the model. Topic models can cluster commonly used words as “topics”¹⁴ and assign each document with proportions of each topic. Every document is then represented by a combination of topics, where the proportions add up to 1. As outputs, we can observe the words that are most representative of each topic and topical proportions. Higher topical proportions mean higher topical prevalence. I choose to run the topic model on the adjectives corpora only because as the entries are paired on the same events and entities, it would be hard to detect distinguishing opinion words. After identifying key adjectives, they can be put into context to be understood.

Figure 4 shows the overall prevalence and words that associate with the seven topics with the highest probabilities. The distinguishing words are either descriptive for events or assertive for evaluations. Figure 5 shows the difference in topic proportions between Chinese Wikipedia text and Baidu Baike text with 95% confidence interval, and each topic is labeled with its representative words. When the difference is positive, the topic is

¹⁴The number of topics are manually chosen. I choose 7 topics after comparing models of less and more topics, because this model could best differentiate topics that are more interpretable.

more prevalent on Chinese Wikipedia. In turn, when the difference is negative, the topic is more prevalent on Baidu Baike. Topic 6, marked by feudal, important, and proactive, is most associated with Baidu Baike. The words in Topic 6 are also frequently used in Marxist Historiography. The other topic that is relatively more prevalent on Baidu Baike is Topic 4, distinguishable by proactive, resolute, and united, which are frequently used by the CCP in meetings as well as during the “new democratic revolution”. Topic 5, represented by big, formal, and successful, is most associated with Chinese Wikipedia.

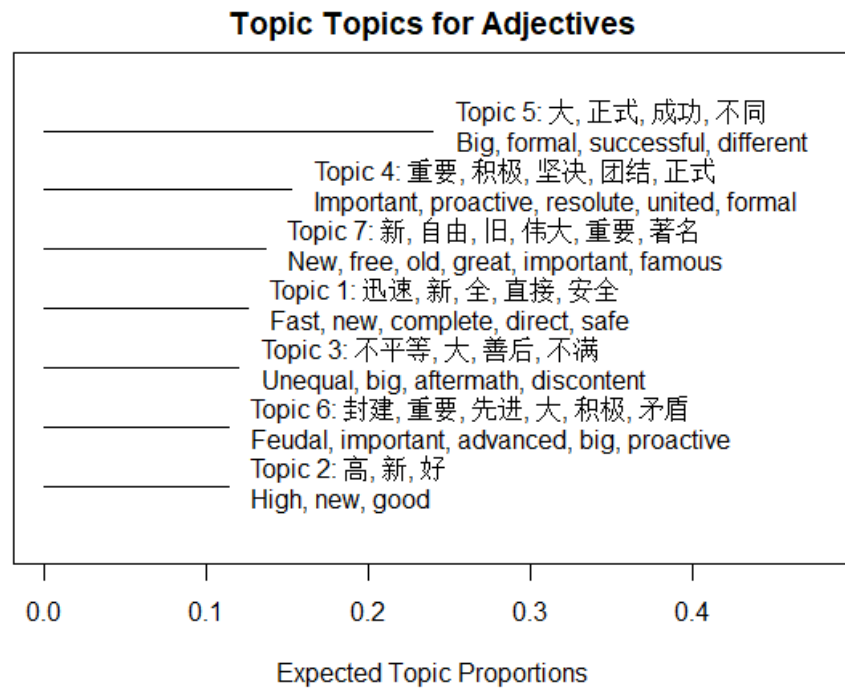


Figure 4: Top Topics for Adjectives

Since the target events and entities in the corpora are the same ones, the different prevalences of usage of adjectives around them show how the events are framed and where subjectivity comes in. To take a closer look at Topic 6, Figure 6 presents two most representative texts in context and the prevalence of the topic over time. The first piece of text is a paragraph that evaluates the land system established by the Taiping Heavenly Kingdom (太平天国) in the Baidu Baike entry of the Land System of the Heavenly Kingdom (天朝田亩制度), which is an equalizing but impractical land policy document by a peasant rebellion power in Late-Qing. The piece comments on the document from the perspective of class struggle and frames it as “revolutionary” against “feudal” control.

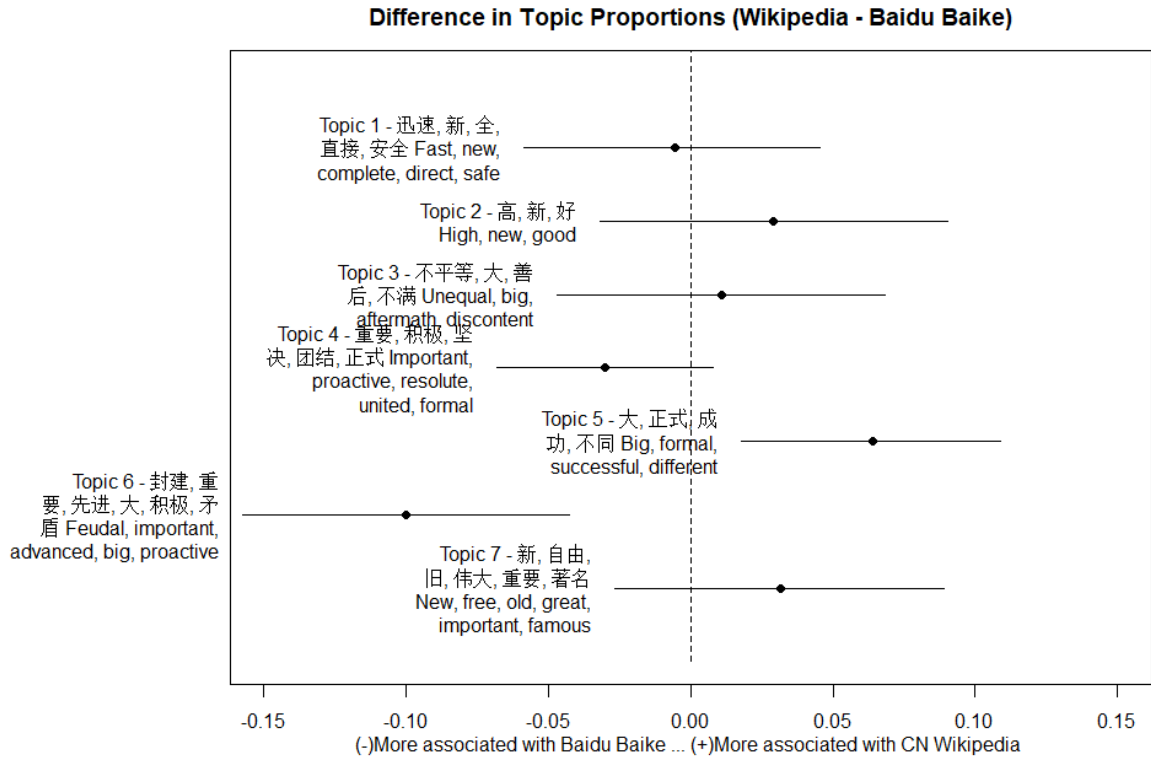


Figure 5: Expected Topical Prevalence Difference between Baidu Baike and Chinese Wikipedia

Frequent use of inflaming adjectives, such as backward (倒退的), reactionary (反动的), and unequivocal (鲜明的), asserts a strong subjective stand over the historical role of the document. The second example is from the Baidu Baike entry of the Three Principles of the People (三民主义) by Sun Yat-sen as political ideology. While the paragraph aims to describe alternative opposing views of the principles, which is welcome by Chinese Wikipedia in providing a whole picture, the perspective also comes from the framework of anti imperialism and feudalism. Instead providing evidence from the opposition side, it rely on empty adjectives, thorough (彻底的), self-contradictory (自相矛盾的), and chaotic (混乱的), to belittle the principles. Putting the strand of adjectives under Marxist Historiography in Topic 6 in the time horizon, they are more prevalent on Baidu Baike throughout the coverage of events and entitles in this study, and are relatively more prominent during the early years of Qing's defeat to foreign powers and the birth period of revolution in early 1900s.

Figure 7 further explores the four topics (Topics 4 through 7) that are more distin-

Topic 6	Topic 6
(«天朝田亩制度»)在资本主义和社会主义时期,它是落后的、倒退的和反动的。但是,在反对封建地主所有制,打破封建大地产对农民的桎梏中,它不仅是一面鲜明的旗帜,而且是进步的 and 革命的,因为它起着推动生产力发展,解放生产力的历史作用。 但是三民主义也存在着历史的局限,主要表现为缺乏明确的、彻底的反帝反封建内容。不支持三民主义者的,他们认为三民主义基本上是东抄西凑,自相矛盾,思想极度混乱的产物。	(Tianchao Tianmu Zhidu, the land system of the Taiping Heavenly Kingdom) In the period of capitalism and socialism, it was behind, backward and reactionary. However, in opposing feudal landlord ownership and breaking the shackles of feudal land ownership to farmers, it is not only a clear banner, but also progressive and revolutionary, because it plays a historical role in promoting and liberating productivity. But the Three Principles of the People also have historical limitations, mainly manifested by the lack of clear and thorough anti-imperialist and anti-feudal content. Those who do not support the Three Principles of the People believe that the Principles are copy-and-pasted and self-contradictory.

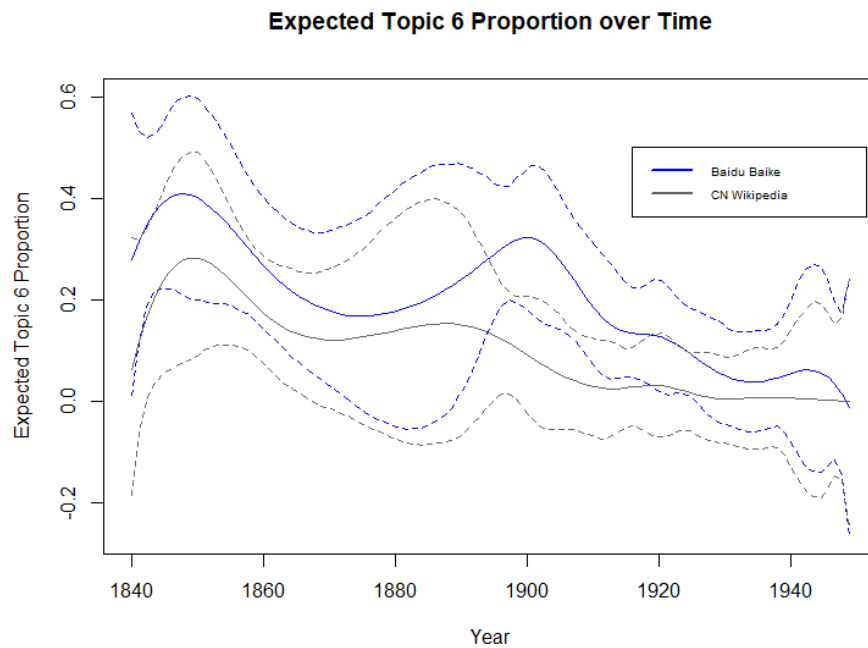


Figure 6: Topic 6 Representative Text and Prevalence over Time

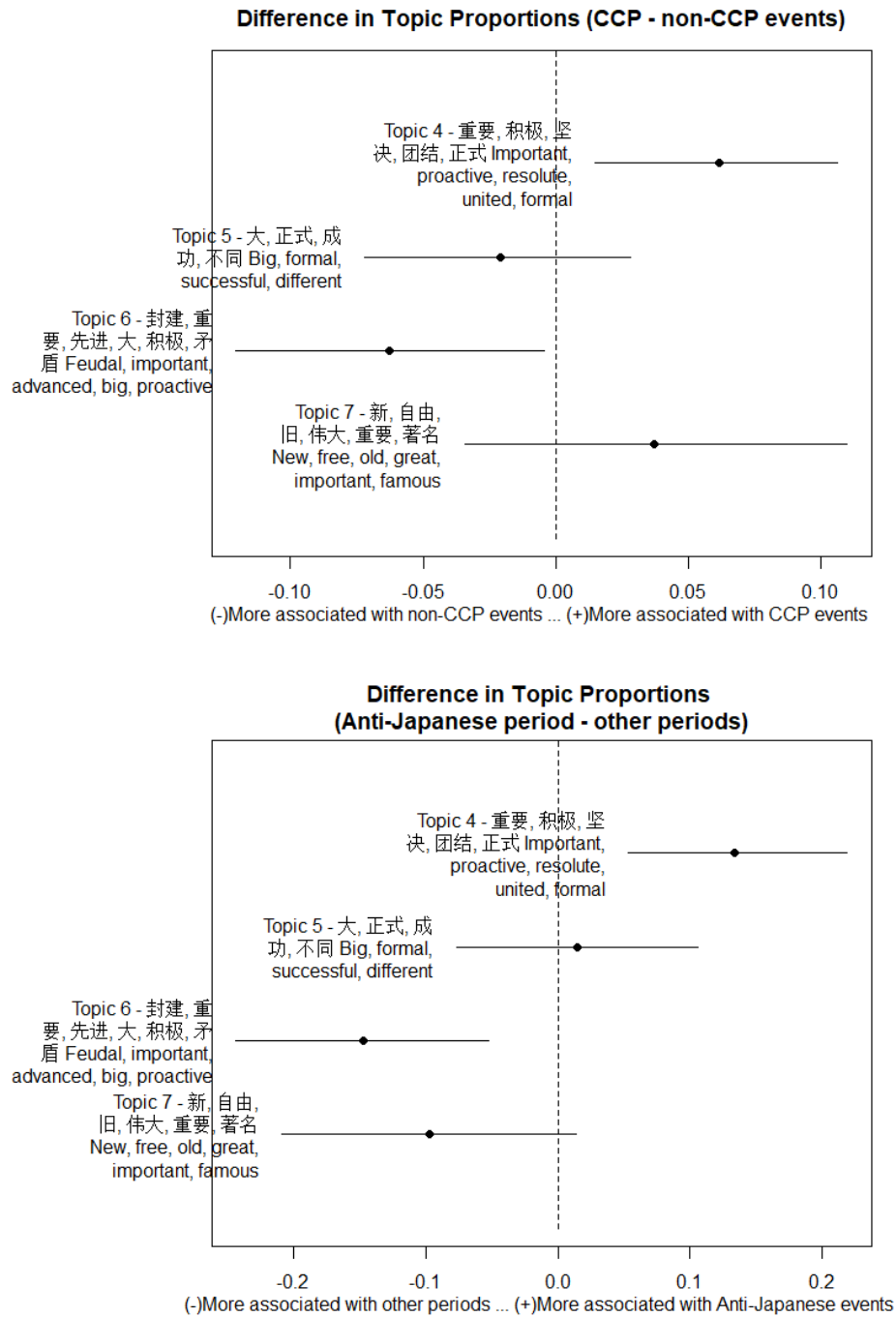


Figure 7: Expected Topical Prevalence Difference over CCP Presence and Anti-Japanese Period

guishable in terms of prevalence on Baidu Baike and Wikipedia. The top panel shows differences in association with CCP presence in events and entities. For the two topics more prevalent on Baidu Baike, Topic 4 is more associated with entities with CCP presence. Topic 6 of Marxist Historiography is less associated with CCP entries, which is in line with the observation from Figure 6. The bottom panel displays expected topic prevalences for the Anti-Japanese period versus other periods. Words of Topic 7, New, Free, Old, Great and Famous, are more likely to be used by Chinese Wikipedia in describing events and entities prior to the Anti-Japanese period.

In this sub-section, I show that Baidu Baike is more subjective than Chinese Wikipedia on the historical narratives in that Baidu Baike uses more adjectives. Examples on Baidu Baike suggest that the adjectives are often helpers for assertive historical judgements and savers for lack of evidence. And subjective assertions on Baidu Baike often come from the official Marxist Historiography. The results also explore which entries are likely to be surrounded by adjectives by topics. Next, I turn to the variations in sentiment of the entries.

4.2 Sentiment and Similarity

Hypothesis 2 expects that Baidu Baike features events and entities where the CCP was a key actor more positively. Moreover, Baidu Baike and Chinese Wikipedia tend to agree on entries on the Anti-Japanese period or on traumatic experiences both in terms of sentiment and general framing. In short, I expect the role of the CCP to be driving the differences while collective traumas to be driving consensus between Baidu Baike and Chinese Wikipedia. I use the average sentiment score as a general measurement and positive to negative word ratio as an indicator for sentiment discrepancy positive-to-negative leverage and examine how they varies across different event features on Baidu Baike and Chinese Wikipedia.

First, Figure 8 plots the predicted the sentiment scores of Baidu Baike and Chinese Wikipedia entries over the four major historical periods.¹⁵ Baidu Baike and Chinese

¹⁵See Columns (1) and (3) of Table 11 in Appendix for the OLS model.

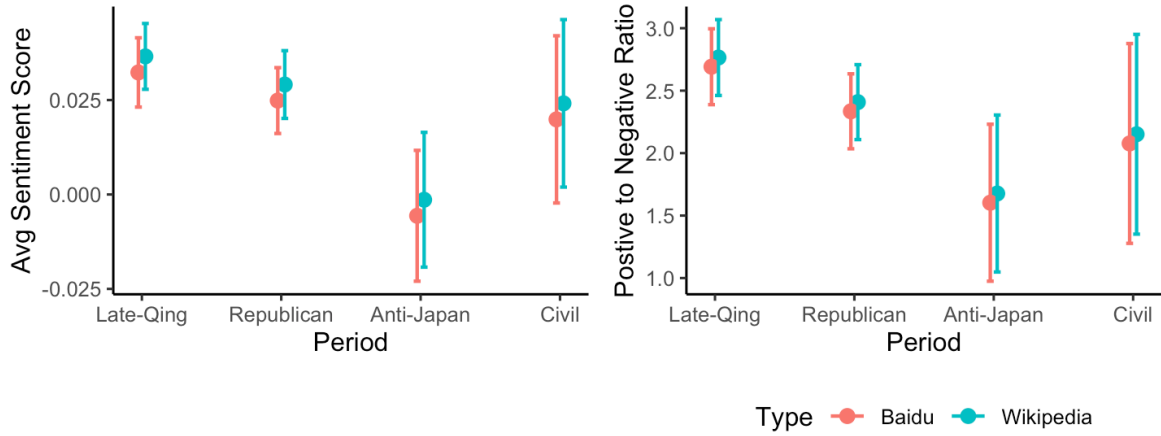


Figure 8: Predicted Values of Average Sentiment Score and Positive to Negative Word Ratio by Period

Wikipedia moves synchronously across periods, which suggests general agreement on the perception over the Century of Humiliation in China. Entries on events during the Anti-Japanese period see a large drop in sentiment scores on both Baidu Baike and Wikipedia, which also have the lowest sentiment scores over all features. Most of the events in the period are battles and wars that led to casualties and losses. They not only are generically surrounded by negative words¹⁶, but also builds a consensus in terms of sentiment in remembering this past, despite discrepancies between Baidu Baike and Chinese Wikipedia.

One preeminent disagreement, however, is on events and entities where the CCP played a key role. In Figure 9,¹⁷ both indicators reveal that Baidu Baike encompass more positive description of events with CCP presence than Chinese Wikipedia. Baidu Baike uses less positive words on events without CCP presence than Chinese Wikipedia. When it comes to events with CCP presence, the positive to negative ratio becomes significantly higher for Baidu Baike. This suggests that the historical role of the CCP is praised more on Baidu Baike, which drives the difference from Chinese Wikipedia.

Finally, in Figure 10,¹⁸ Baidu Baike's account of traumatic events are more negative

¹⁶The most negative words in ANTUSD that appear in this corpora are casualty and war related, see Appendix.

¹⁷See Columns (2) and (4) of Table 11 in Appendix for the OLS model with interaction.

¹⁸See Columns (3) and (6) of Table 11 in Appendix for the OLS model with interaction.

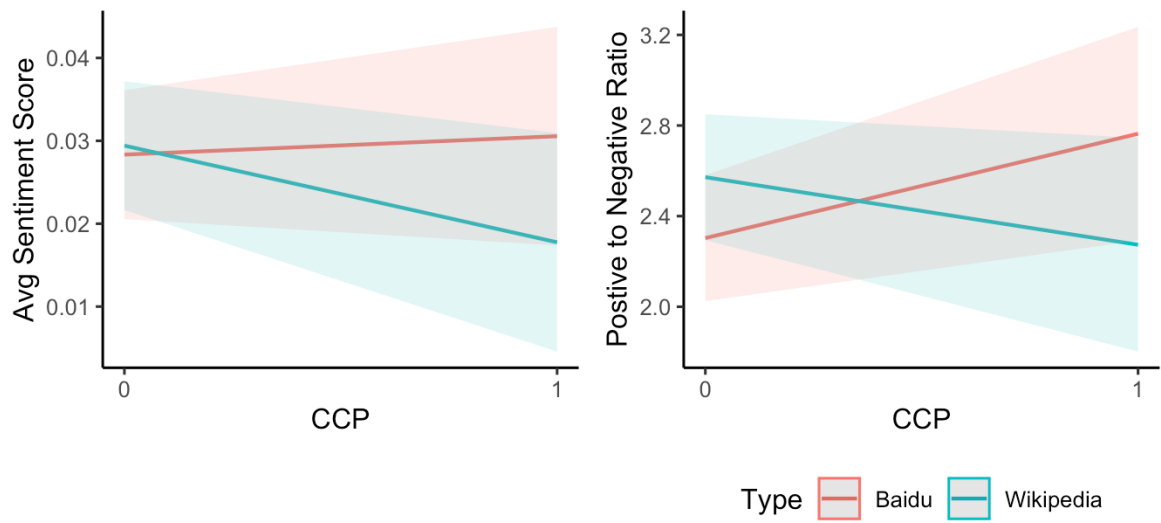


Figure 9: Predicted Values of Average Sentiment Score and Positive to Negative Word Ratio, CCP Presence

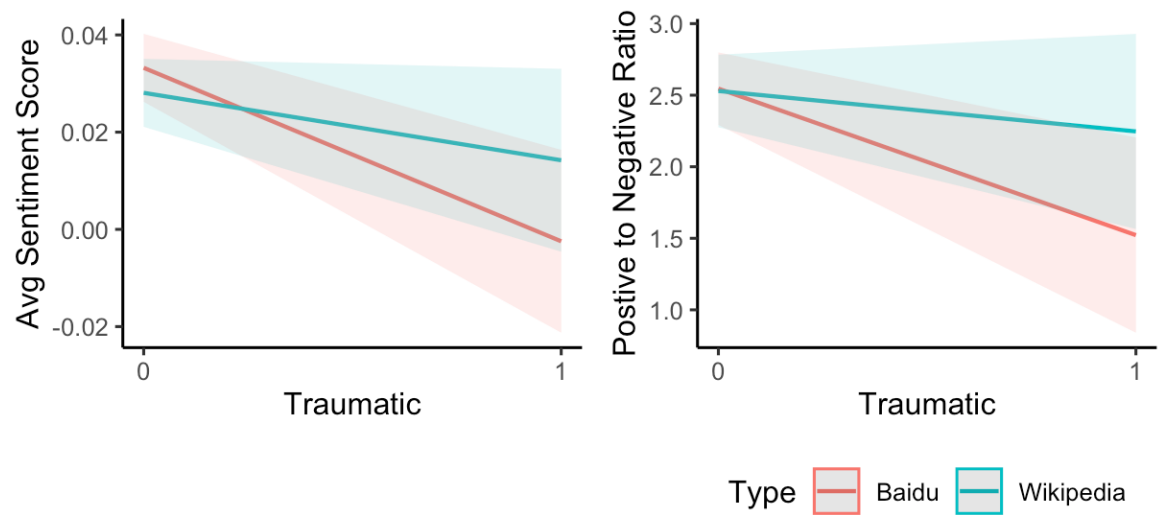


Figure 10: Predicted Values of Average Sentiment Score and Positive to Negative Word Ratio, Traumaticity

than Chinese Wikipedia. The marginal negative effect of traumatic events is smaller for Chinese Wikipedia than for Baidu Baike. Nevertheless, the differences in the positive to negative word ratio and average sentiment scores are relatively small in absolute values. Figure 11 looks at the variations in depicting traumatic events from text similarity. Higher cosine similarity score indicates that the pair of entries on the same event or entity are more similar. The scatter plot shows there are three main waves of traumas: the beginning of foreign interventions in the 1840s, the defeat to Japan in the 1890s, and the phase of Anti-Japanese War. On average, a set of paired entries on the same event on Baidu Baike and Wikipedia is more similar in terms of words used if the event is traumatic, i.e. involving territory and sovereignty loss, defeats, and casualties on the Chinese side.

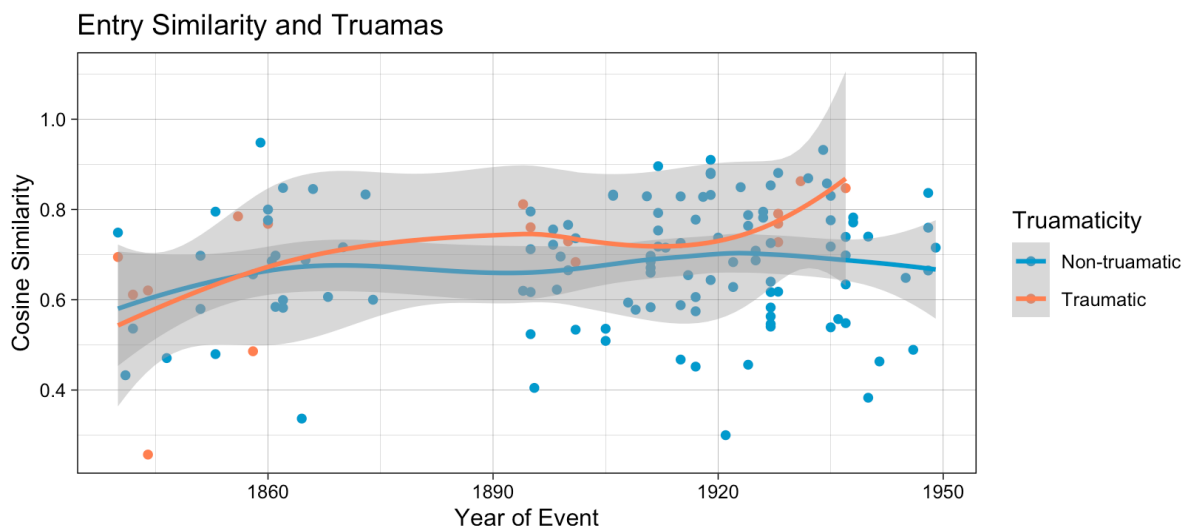


Figure 11: Distribution of Entry Similarity by Year, Events

The exploration in terms of sentiment and text similarity provides evidence on substantive drivers behind agreements and disagreements between Baidu Baike and Chinese Wikipedia, despite information control mechanisms. Baidu Baike is found to frame the role of the CCP in Modern Chinese history more positive, diverging its information away from Chinese Wikipedia. Common chosen traumas possess the potential for Baidu Baike and Chinese Wikipedia to overcome differences, which also reflects the power of the Century of Humiliation narrative.

4.3 Description: Who are the editors and how do they edit?

Platform-specific accessible user profiles make different features of Baidu Baike and Wikipedia users available. In this sub-section, I supplement the discussion on information and historical knowledge negotiation by looking at features of the information generators. The population of users here are those who have edited at least one of the entries studied in this paper either on Baidu Baike or Chinese Wikipedia. As is mentioned in Section 2, all Baidu Baike users have their accounts, while Chinese Wikipedia users include anonymous IP addresses.

A median Baidu Baike user in this population have gained 259 wealth points which is approximately equivalent to 4.32 RMB of worth in Baidu Online Mall. She would own a Baidu level-3 account with a Baidu Baike editing approval rate of 84%, which is below the threshold for many entries: level-4 account and a 85% approval rate. For Chinese Wikipedia user population, 2% of them have administrative rights, 1% of them are blocked for disruptive editing, and 63% of them did their edits anonymously.

Table 6: One-sided T-Tests for Entry Level Editing Activities

	estimate	Wikipedia mean	Baike mean	t statistic	p.value
Editor Lifetime Edits	5333.13	7183.92	1850.79	6.82	0.00
Entry Edits	352.22	535.01	182.79	5.26	0.00
Entry Editors	66.83	190.84	124.02	2.65	0.00
Entry Editor/Edit	0.21	0.47	0.67	11.2	0.00
Entries Edited in Corpora	0.46	1.88	1.42	10.52	0.00

Table 6 compares editing activities on Baidu Baike and Chinese Wikipedia on entries in the corpora. On average, there are higher numbers of edits and more editors involved on Chinese Wikipedia than on Baidu Baike ($p < 0.01$). Wikipedia also shows a lower editor to edit ratio than Baidu Baike, which indicates that Baidu Baike editors tend to make one-time unique edits. As is discussed in 4.1, Chinese Wikipedia is already less subjective despite the expected neutralizing impact of more unique contributors. The moderate editor to edit ratios therefore also signal editing activities on Chinese Wikipedia

are more interactive.

Table 7: One-sided T-Tests for User Level Lifetime Edits

	estimate	Wikipedia mean	Baike mean	t statistic	p.value
(1)	5333.13	7183.92	1850.79	6.82	0.00
(2)	585.21	2435.78	1850.57	1.97	0.02
(3)	4992.03	7183.92	2191.89	6.33	0.00
(4)	243.90	2435.78	2191.89	0.78	0.21

(1): Wikipedia registered users edits (displayed) vs. Baidu Baike users approved edits (displayed)

(2): All Wikipedia users edits (displayed) vs. Baidu Baike users approved edits (displayed)

(3): Wikipedia registered users edits (displayed) vs. Baidu Baike users estimated total attempted edits

(4): All Wikipedia users edits (displayed) vs. Baidu Baike users estimated total attempted edits

Since the Baidu Baike user profile page show versions of edits the user submitted that have been approved and her lifetime approval rate, I estimate the total number of edits for each user in my population. For each Baidu Baike user, I have both the number of edits displayed and the number of edits attempted. For Wikipedia, the user profile extraction API only provides lifetime edits for registered users, not anonymous IP addresses. I calculate numbers of edits occurred in my entry population by IP addresses to substitute for missing values of anonymous users.

Table 7 shows comparisons among those four groups of edit numbers data. Rows (1) and (2) indicate that for the user populations editing the 132 modern Chinese historical events, the number of approved and displayed edits by Baidu Baike users are significantly lower than the number of edits by Wikipedia users, both registered and all including anonymous ones. In Row (4), however, when comparing estimated total attempted edits by Baidu Baike users to all Wikipedia users, the number is no longer statistically different. If anything, this suggests the underlying vibrancy of editing activities by users are no different across Baidu Baike and Chinese Wikipedia. In other words, the differences in displayed historical knowledge is likely strongly influenced by Baidu’s control mechanisms.

5 Discussion and Conclusion

In this paper, I propose to explain the differences in historical information and knowledge displayed on Baidu Baike and Chinese Wikipedia with a framework of equal and unequal information negotiation under different exposure to censorship. Not surprisingly, Baidu Baike is more subjective in writing the Century of Humiliation than Chinese Wikipedia, employing Marxist historiography and in favor of the CCP. I also document partial evidence that the editing activities are not different between Baidu Baike and Chinese Wikipedia, which emphasizes the strong role of Baidu's ex ante oversight as the edits displayed are very different. Despite objective forces driving the differences in historical information and knowledge on Baidu Baike and Chinese Wikipedia, common traumatic experience stands out as a substantive internal force converging the narratives on events of traumas.

However, this is only an initial effort and there is also much room for improvement. First, the dictionary based text analysis methods and unsupervised topic modeling are less accurate than desired. A more customized dictionary and more human coding efforts are required to improve the quality of the measurements. Another effort can be made to enlarge the corpora of interest and identify coverage differences of Chinese Wikipedia and Baidu Baike. Since Baidu Baike and Chinese Wikipedia both have their historical versions of entries available, it is also worth tracing the major editorial paths and constructing measurements from text over time to study the effect of the Great Firewall on the entries.

To be sure, it is natural for collective memories to have national perspectives and citizens tend to overestimate their own countries' historical position in general. People from 11 countries respectively overclaimed their responsibility in the World War II and showed national narcissism and self-centeredness (Roediger et al., 2019). Huang and Liu (2018) finds that Chinese citizens tend to overestimate China's historical achievements and those who do overestimate have higher levels of national identity. Crowdsourcing encyclopedia platforms offer an opportunity to study relative positioning of historical narrative and collective memory as well as their political implications.

The quality of historical knowledge as is presented on online encyclopedias have im-

plications for information consumption. Pan and Roberts (2019) have shown that the block of Wikipedia reduced accidental exposure to historical information of people from mainland China. Without the presence of the largest Chinese speaking population in the world, global knowledge bases as is represented by Wikipedia is also diminished. With little to no alternative sources of explanations for the majority in China, under a potentially consolidated inter-generation trauma of the Century of Humiliation, the West can be easily framed as the enemy for political purposes. “New enemies involved in current conflicts may be perceived as extensions of an old enemy from a historical event.”(Volkan, 1997). The mentality risen from “chosen trauma” and “chosen glory” is therefore worthy of more attention.

References

- Anderson, Benedict. 2006. Imagined communities: Reflections on the origin and spread of nationalism. Verso books.
- Assmann, Jan and John Czaplicka. 1995. “Collective memory and cultural identity.” New german critique (65):125–133.
- Cantoni, Davide, Yuyu Chen, David Y Yang, Noam Yuchtman and Y Jane Zhang. 2017. “Curriculum and ideology.” Journal of Political Economy 125(2):338–392.
- Coble, Parks M. 2007. “China’s “new remembering” of the Anti-Japanese War of Resistance, 1937–1945.” The China Quarterly 190:394–410.
- Cohen, Paul A. 2002. “Remembering and forgetting national humiliation in twentieth-century China.” Twentieth-Century China 27(2):1–39.
- Halbwachs, Maurice et al. 1992. On collective memory. University of Chicago Press.
- Hobbs, William R and Margaret E Roberts. 2018. “How sudden censorship can increase access to information.” American Political Science Review 112(3):621–636.
- Huang, Haifeng and Xinsheng Liu. 2018. “Historical knowledge and national identity: Evidence from China.” Research & Politics 5(3):2053168018794352.
- Jović, Dejan. 2004. ““Official memories’ in post-authoritarianism: an analytical framework.” Journal of southern Europe and the Balkans 6(2):97–108.
- Kaufman, Alison Adcock. 2010. “The “century of humiliation,” then and now: Chinese perceptions of the international order.” Pacific Focus 25(1):1–33.
- Lam, Onyi and Eddie Lin. 2017. “A Computational Analysis of History Textbooks in Mainland China, Hong Kong and Taiwan.” Unpublished Manuscript.

- Miller, Blake Andrew Phillip. 2018. "The limits of commercialized censorship in China."
- Ng, Jason Q. 2013. "Who's the Boss? The difficulties of identifying censorship in an environment with distributed oversight: A large-scale comparison of Wikipedia China with Hudong and Baidu Baike." The Citizen Lab . Accessed at <https://citizenlab.ca/2013/08/a-large-scale-comparison-of-wikipedia-china-with-hudong-and-baidu-baike/>.
- Pan, Jennifer and Margaret E. Roberts. 2019. "Censorship's Effect on Incidental Exposure to Information: Evidence from Wikipedia." SAGE Open .
- Pentzold, Christian. 2009. "Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place." Memory studies 2(2):255–272.
- Roediger, Henry L, Magdalena Abel, Sharda Umanath, Ruth A Shaffer, Beth Fairfield, Masanobu Takahashi and James V Wertsch. 2019. "Competing national memories of World War II." Proceedings of the National Academy of Sciences 116(34):16678–16686.
- Shin, Gi-Wook and Daniel C Sneider. 2011. History textbooks and the wars in Asia: Divided memories. Vol. 31 Routledge.
- Stockmann, Daniela. 2012. Media Commercialization and Authoritarian Rule in China. Cambridge: Cambridge University Press.
- Volkan, Vamik. 1997. Bloodlines: From Ethnic Pride to Ethnic Terrorism. Farrar, Straus and Giroux.
- Wang, Shih-Ming and Lun-Wei Ku. 2016. ANTUSD: A large Chinese sentiment dictionary. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2697–2702.
- Wang, Zheng. 2014. Never forget national humiliation: Historical memory in Chinese politics and foreign relations. Columbia University Press.
- Zhang, Xiaoquan Michael and Feng Zhu. 2011. "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia." American Economic Review 101(4):1601–15.

Appendix

List of Entry Names

Table 8: List of Entries

Baidu Baike Entry	CN Wikipedia Entry	Baidu Baike Entry	CN Wikipedia Entry	Baidu Baike Entry	CN Wikipedia Entry
第一次鸦片战争	第一次鸦片战争	义和团运动	义和团运动	北洋水师	北洋水師
洋务运动	洋务运动	中国同盟会	中国同盟会	天津教案	天津教案
南京条约	南京条约	预备立宪	立憲運動	马嘉理案	馬嘉理事件
太平天国运动	太平天国	黄花岗起义	黃花崗起義	中学为体，西学为用	中体西用
第二次鸦片战争	第二次鸦片战争	保路运动	保路运动	康有为	康有为
甲午战争	甲午战争	武昌起义	武昌起义	梁启超	梁启超
八国联军侵华	八国联军之役	二次革命	二次革命	清末新政	清末新政
辛丑条约	辛丑条约	新青年	新青年	严复	嚴復
辛亥革命	辛亥革命	新文化运动	新文化运动	孙中山	孫中山
中华民国临时约法	中華民國臨時約法	新生活运动	新生活運動	三民主义	三民主義
五四运动	五四运动	护国运动	護國戰爭	黄埔军校	黃埔军校
南昌起义	南昌起义	袁世凯复辟	中华帝国	保路运动	保路运动
八七会议	八七会议	护法运动	護法運動	申报	申報
秋收起义	秋收起义	天朝田亩制度	天朝田亩制度	黎元洪	黎元洪
广州起义	广州起义 (1927 年)	资政新篇	资政新篇	袁世凯	袁世凱
井冈山会师	朱毛会师	三民主义	三民主义	东南互保	東南互保
东北易帜	东北易帜	民报	民报	曹锟	曹錕
九一八事变	九一八事变	七一五政变	七一五事变	滦州起义	滦州兵变
一二八事变	一·二八事變	中共三大	中国共产党第三次全国代表大会	冯玉祥	馮玉祥
遵义会议	遵义会议	国共合作	聯俄容共	直系军阀	直系
一二九运动	一二·九运动	五卅运动	五卅运动	吴佩孚	吳佩孚
西安事变	西安事变	北伐战争	國民革命軍北伐	孙传芳	孫傳芳
卢沟桥事变	七七事变	四一二政变	四一二事件	段祺瑞	段祺瑞
八一三事变	八一三事变	淞沪会战	淞沪会战	丁巳复辟	張勳復辟
平型关大捷	平型关战役	林则徐	林则徐	奉系军阀	奉系
南京大屠杀	南京大屠杀	海国图志	海国图志	张作霖	張作霖
台儿庄战役	台兒莊戰役	耆英	耆英	张学良	張學良
百团大战	百团大战	魏源	魏源 (清朝)	皖系军阀	皖系
皖南事变	皖南事变	璦琿条约	璦琿條約	南北和谈	1919 年中國南北和平談判
双十协定	雙十協定	洪秀全	洪秀全	直皖战争	直皖戰爭
辽沈战役	辽沈战役	曾国藩	曾国藩	傅作义	傅作义
淮海战役	淮海战役	湘军	湘军	直奉战争	直奉戰爭
平津战役	平津战役	左宗棠	左宗棠	济南惨案	五三慘案
渡江战役	渡江战役	李鸿章	李鴻章	皇姑屯事件	皇姑屯事件
望厦条约	望廈條約	张之洞	张之洞	阎锡山	閻錫山
黄埔条约	黃埔條約	捻军起义	捻军	陈独秀	陳獨秀
金田起义	金田起义	云南回民起义	云南回变	蔡元培	蔡元培
天津条约	天津条约	同治中兴	同光中兴	胡适	胡適
辛酉政变	辛酉政变	慈禧	慈禧太后	李大钊	李大釗
北京条约	北京条约	总理各国事务衙门	總理各國事務衙門	鲁迅	魯迅
兴中会	兴中会	京师同文馆	京師同文館	张国焘	張國焘
马关条约	马关条约	爱新觉罗·奕訢	奕訢	中国共产党第一次全国代表大会	中国共产党第一次全国代表大会
戊戌变法	戊戌变法	江南机器制造总局	江南機器製造總局	第二次国共合作	國共關係
长征	长征	轮船招商局	輪船招商局	汪精卫	汪精卫

Top Positive and Negative Words in Corpora

Table 9: Top 20 Positive and Negative Words in the Corpora, ANTUSD

Positive Words	Sentiment Score	Negative Words	Sentiment Score
喜爱	0.8875107	杀死	-0.9125600
喜欢	0.8750000	死伤	-0.9125600
欢喜	0.8750000	伤害	-0.9000000
喜好	0.8426862	击落	-0.9000000
安适	0.8343103	失落	-0.9000000
爱好	0.8301969	暴乱	-0.9000000
精美	0.8274877	杀害	-0.9000000
精明	0.8188178	破坏	-0.9000000
正好	0.8176862	刺死	-0.8875600
美好	0.8176862	惨死	-0.8875600
爱好和平	0.8149924	低落	-0.8750000
敬爱	0.8125107	击倒	-0.8750000
平安	0.8090983	刺伤	-0.8750000
未失	0.8000000	压倒	-0.8750000
温和	0.8000000	压碎	-0.8750000
真正	0.8000000	反倒	-0.8750000
和平	0.7997880	失掉	-0.8750000
热爱	0.7991829	恶毒	-0.8750000
清明	0.7945070	杀掉	-0.8750000
不死	0.7930274	毒害	-0.8750000

Comparison Group of the Sentiment Measurement

Table 10: Comparison Group and Average Sentiment Scores

	Figure Entries	Baidu Baike Mean	CN Wikipedia Mean
Positive	毛泽东, 江泽民, 胡锦涛, 习近平, 周恩来, 朱镕基, 温家宝, 李克强, 邓小平, 曾庆红, 华国锋, 李鹏, 杨尚昆, 谷牧, 吴邦国, 李岚清, 纪登奎, 乔石, 邹家华, 李瑞环, 俞正声, 张高丽, 田纪云, 回良玉, 李源潮, 贾庆林, 姚依林, 张立昌, 尉健行, 姜春云, 李铁映, 王兆国, 罗干, 刘靖基, 杨汝岱, 王光英, 彭佩云, 刘云山, 丁关根, 彭真, 胡启立, 曾培炎	0.10	0.06
Negative	林彪, 王洪文, 张春桥, 江青, 姚文元, 刘晓波, 丹增嘉措, 陈水扁, 黄之锋, 黎智英	-0.002	0.02

OLS Results for Sentiment Analysis

Table 11: OLS Models

	<i>Dependent variable:</i>					
	Average Sentiment Score			Positive to Negative Ratio		
	(1)	(2)	(3)	(4)	(5)	(6)
Wikipedia	0.004 (0.006)	0.009 (0.006)	0.004 (0.006)	0.427** (0.205)	0.634*** (0.226)	0.394* (0.211)
Republican	-0.007 (0.005)			-0.316* (0.180)		
Anti-Japan	-0.038*** (0.009)			-0.971*** (0.332)		
Civil War	-0.012 (0.012)			-0.569 (0.414)		
CCP		0.001 (0.008)			0.416 (0.275)	
Wikipedia*CCP		-0.011 (0.011)			-0.679* (0.390)	
Truamatic			-0.038*** (0.010)			-1.133*** (0.364)
Wikipedia*Truamatic			0.025* (0.014)			0.910* (0.515)
Edit	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	-0.0002 (0.0002)	-0.0002 (0.0002)	-0.0002 (0.0002)
Editor/Edit	0.034** (0.016)	0.038** (0.017)	0.046*** (0.016)	1.316** (0.587)	1.439** (0.591)	1.700*** (0.583)
Constant	0.013 (0.013)	0.003 (0.012)	0.002 (0.012)	1.809*** (0.450)	1.380*** (0.432)	1.443*** (0.422)
Observations	264	264	264	264	264	264
R ²	0.087	0.032	0.083	0.082	0.057	0.082
Adjusted R ²	0.065	0.013	0.065	0.061	0.039	0.064
Residual Std. Error	0.038 (df = 257)	0.039 (df = 258)	0.038 (df = 258)	1.362 (df = 257)	1.377 (df = 258)	1.360 (df = 258)
F Statistic	4.070*** (df = 6; 257)	1.708 (df = 5; 258)	4.644*** (df = 5; 258)	3.836*** (df = 6; 257)	3.144*** (df = 5; 258)	4.582*** (df = 5; 258)

Note:

*p<0.1; **p<0.05; ***p<0.01