CASSANDRA CASE STUDY

Hong Kong: Longitudinal and Synchronic Characterizations of Protest News between 1998 and 2020

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING
CENTER FOR LANGUAGE AND SPEECH PROCESSING
apsa
Special Projects Fund

CASSANDRA

CASSANDRA is an academic initiative that brings together computational science and social science to increase visibility, scholarship, and communication between the two fields, while at the same time addressing the challenges of conducting valid, open, and ethical research at the nexus of political and computational sciences.

CASSANDRA is managed through the Center for Language and Speech Processing of the Johns Hopkins University, Whiting School of Engineering, and funded by a grant from the APSA Special Project Fund.

Abstract

This case study builds on McCarthy and Dore's *Hong Kong: Longitudinal and Synchronic Characterizations of Protest News between 1998 and 2020* (2022), investigates longitudinal and synchronic characterizations of news protests in Hong Kong between 1998 and 2020 through the application of natural language processing to its 4522 articles and thereby study patterns of journalistic practice across newspapers. This case study sheds light on whether depth and/or manner of reporting changed over time, and if so, in what ways, or in response to what. In its focus and methodology, this paper helps bridge the gap between validity-focused methodological debates and the use of computational methods of analysis in the social sciences.

1. Introduction

Protests constitute an important means in contemporary societies by which citizens voice their concerns. However, protests' ability to communicate their messages and achieve their outcomes depends significantly on whether and how mass media, and especially news- papers, portray them (Agnone, 2007; King, 2014). Newspapers, in fact, can either amplify and legitimize the protesters' voices (Gamson and Wolfsfeld, 1993) or marginalize and delegitimize protests by portraying them as dangerous or irrelevant (Boykoff, 2006; Solomon, 1996). To investigate the dynamic relationship between protests and newspapers, we have constructed the *Hong Kong Protest News Dataset*, an original collection of news articles covering the occurrence and evolution of protests in Hong Kong between 1998 and 2020.

Protest participation has long been an undercurrent in Hong Kong's political culture (Rawnsley and Rawns- ley, 2002) dating back to British colonial rule, and has evolved from the bloody riots of the 1960s to the protests of 2019–2020, when up to two million people took to the streets against the proposed amendment to the Fugitive Offenders and Mutual Legal Assistance in Criminal Matters Legislation Bill (ELAB). Hong Kong protests captured the world's attention with defiant crowds commemorating the 1989 Tiananmen Square incidents, nostalgically marking the transfer of sovereignty from the UK back to China every July 1 since 1997, and students blockading roads for 79 days in the Admiralty district during the pro-democracy Occupy Central protests in 2014 (Weiss and Aspinall, 2012). The news value of these actions grew, as early demonstrations to voice dissent morphed into an increasingly violent anti-government, anti-Beijing movement with demands for greater democracy.

We apply natural language processing (NLP) to study patterns of journalistic practice across newspapers, shed light on whether depth and/or manner of reporting changed over time, and if so, in what ways, or in response to what. As language is at the heart of our research, NLP emerges as especially important for its ability "to analyze signals ranging from simple lexical clues to word clusters to choices of syn- tactic structure" (Boydstun et al., 2014, 2) as well as its speed, scale, reliability, and granularity when analyzing text. This case study builds on (Scharf et al., 2021) who consider a subset of our techniques with a preliminary form of the dataset, and (McCarthy et al., 2021) who focus solely on the recent anti-ELAB protests. It also complements a small collection of articles, currently under review and/or being drafted, that aims at bridging the gap between "validity-focused methodological debates" (Baden et al., 2021, 13) and the use of computational methods of analysis in the social sciences.

2. The Hong Kong Protest News Dataset

As newspapers' coverage remains one of the most useful records of protest events (Earl et al., 2004), we took steps to include a diverse array of news sources, even though Chinese, including Hong Kong-SAR, North American and British newspapers sit at opposite ends of the spectrum in terms of ownership and state control. The corpus of articles we construct comes from six western-based, English language newspapers: The New York Times (NYT), The Wall Street Journal (WSJ), The Washington Post (WaPo), The Financial Times (FT), The Guardian, and The Times; and two Hong Kong–based, English language newspapers: China Daily and South China Morning Post (SCMP).

Current limitations of natural language processing (NLP) comparison across languages led us to include only English-language newspapers (Earl et al., 2004; Baden et al., 2021). While multilingual NLP approaches are being proposed, they are typically proven only for languages related to English (Mimno et al., 2009; Reber, 2019; Chan et al., 2020). In fact, Bender (2009) discusses the limitations generally, and Baden et al. (2021) focus on specific issues in multilingual NLP for the social sciences. Moreover, a recent WMT shared task (Barrault et al., 2019) shows why translation as a preprocessing step is sub-optimal (cf. Field et al., 2018), as current translation systems struggle to maintain coherence across long document contexts. Besides these NLP limitations, other factors validate our methodological choice. The SCMP and China Daily are the printed English-language newspapers with the largest readership in mainland China and Hong Kong. Finally, our methodological approach is consistent with the work of Bhatia (2015), Yu (2015); Wong and Liu (2018); Du et al. (2018), Lee (2014), McCarthy and Dore (2022), and Scharf et al. (2021), who rely on English language news sources to investigate news portrayals of protests in Hong Kong.

We use news articles focusing on eight non-randomly selected episodes of civic unrest in Hong Kong to com- pare their news value and newsworthiness in the volatile social and political setting of post-handover Hong Kong (Chan and Lee, 1984; Lee, 2014; Tsfati and Walter, 2019). We focus on (i) the 1998–2002 July anniversary marches; (ii) the 2003 protests against national security reform; (iii) the 2004–2019 July 1 protests; (iv) the 2006–07 save the Star Ferry Pier protests; (v) the 2012 Protests against Moral and National Education; (vi) the 2014 Occupy Central protests; (vii) the 2016 Riots; and (viii) the 2019–2020 anti-extradition protests. Taken together, these protests represent a sustained and organized citizens' effort asking Hong Kong and Chinese authorities for a clear and faster path to democratization for Hong Kong (Chan, 2015; Wong, 2021) and as such have received significant coverage in both Hong Kong– and western-based newspapers.

The articles were collected through keyword-based searches in ProQuest Newspapers for the western English-language newspapers, and Newsbank Access World News Research Collection for the English- language Hong Kong newspapers. We searched for the keywords "Hong Kong" + "protests", "Hong Kong" + "rallies", "Hong Kong" + "marches", and "Hong Kong" + "riots".

We used the East Coast edition for the NYT and WSJ, the UK edition for the FT, The Guardian, and The Times, and Hong Kong edition for China Daily. To be eligible for collection, articles had to be at least 300 words long and to focus on the protests. A one-by-one, manual screening process eliminated irrelevant items such as eventual duplicates within each publication, readers' letters, and (crucially) articles that included any of the chosen keywords but whose content was not relate to the Hong Kong protest incidents. Following the manual screening, we retained a total of 4522 articles; 793 articles come from western-based newspapers and 3729 from Hong Kong–based newspapers, with a mean length of 783 tokens.

The Hong Kong Protest News Dataset has a 22-year time horizon, spanning from January 1, 1998, to June 30, 2020, to capture changes in how Hong Kong– and western-based news sources cover protests in Hong Kong, and test the relevance and robustness of changes in how newspapers treat protests over time. The ex- tended time horizon together with the size of our sample represent a significant departure from other datasets on Hong Kong protests, which tend to include a much smaller samples of articles, focus on a particular episode of protest, or attempt comparisons between no more than two incidents of protests at different points in time. For instance, Bhatia (2015) uses approximately 100 articles the SCMP published over the last two months of the 2014 Occupy Central protests to understand the SCMP's characterization of those protests. Yu (2015) uses 249 news stories to examine the frames that the SCMP, the NYT, and The Guardian use in their coverage of the 2014 Occupy Central protests. Wong and Liu (2018) examine newspapers' representations of the aggressive behavior of social actors in the 2014 Occupy Central protests based on 875 articles from the China Daily and the SCMP. Du et al. (2018) rely on 191 articles from the FT, the NYT, Ming Pao, People's Daily, and the United Daily News to show how differently these newspapers frame news stories about the 2014 Occupy Central protests. Lee (2014) uses 1,767 articles from the Apple Daily, the Oriental, and Ming Pao to investigate whether news organizations exercise any social control function in their discussion of protests that took place in Hong Kong between 2001 and 2012.

3. Methods

In this research, we conceive of interdisciplinarity as the space between computer science and social science. An inquiry must go farther than matching computational techniques to a social science research question (O'Connor et al., 2011); it involves the design of a synergistic methodology that connects the norms and standards for empirical evidence from these two strange bedfellows.

This means partnering computer science's preference for the structured, generalizable, and objective with the unstructured, critical, and contextual that the social sciences champion. This level of interdisciplinarity allows moving beyond individual findings to explanations of their broader importance and contextual understanding. Skepticism can remain toward findings not drawn from the standard practices of one's own field (Armstrong, 1967). To assuage doubts, we leverage *predictive validity* —i.e., expected correspondence between a measure and exogenous events uninvolved in the measurement process -- and *convergent validity* — i.e., correlation with other measures of the same construct" (Quinn et al., 2010; Grimmer and Stewart, 2013).

To meet the challenges inherent in operationalizing our interdisciplinary research, we use a mixed-method approach. As language is at the heart of our research, computational analysis emerges as important for its ability "to analyze signals ranging from simple lexical clues to word clusters to choices of syntactic structure" (Boydstun et al., 2014) as well as its speed, scale, and granularity. In using statistical techniques to analyze text, the case study builds on research of Field et al. (2018) on the use of word embedding similarity, topic models, and dependency parsing to generate clues toward differing portrayals of race and gender in US history textbooks; Field et al. (2018), who relate the content of Russian state-run news articles to the nation's economic performance to push an agenda of distraction; Mosteller and Wallace (1984) and Bergsma et al. (2012) on content analysis and stylometry in consideration of authorship; Jatowt and Duh (2014) and Kulkarni et al. (2015) on the detection of shifts in word meaning, whether between groups or over time, and gradual or at specific change points; Wijaya and Yeniterzi (2011) on whether relevant historical events lead to differences in word usage; and Blei and Lafferty (2007) on tracking shifts in entire topics, rather than in the semantics of individual words.

In terms of qualitative analysis, we employ qualitative content analysis (Fulcher, 2010; Howitt and Cramer, 2007; Braun and Clarke, 2006) and descriptive interference—in the sense of King et al. (1995)—to corroborate the results emerging from the application of computational techniques to BOLD. Our choice to use qualitative analysis to complement quantitative techniques builds on Achen and Snidal (1989)'s recommendation to use historical case studies as a useful complement to statistical research; their

plea was strengthened by Verba's work in the early 1990s (Verba et al., 1993, 1995; Verba, 1996), and Tarrow (1995), which openly called for bridging qualitative and quantitative modes of research in social science. In the last decade, many authors have made a quantum leap Levy (2007) in social science methodology by providing a highly structured approach to qualitative analysis (Coppedge, 1999; Gerring, 2004; Lieberman, 2010).[1]

3.1 Topic Modeling

Cohen states that "newspapers may not be successful in telling people what to think, but they are stunningly successful in telling readers what to think about" (Cohen, 1963, 13). Newspapers do so through agenda-setting (McCombs and Shaw, 1972), which steers perceptions of issue importance, and framing (Entman, 2006), which involves selection and guides how issues should be interpreted. The interplay of agenda setting, and framing provides a unique vantage point for readers learn about a given issue. We tested for which frames Hong Kong- and western-based newspapers used to characterize the protests in Hong Kong protests.

We build on established scholarship and use topic models to find the frames used in the articles across news sources and over time (Jacobi et al., 2016; Dehler-Holland et al., 2021; Ylä-Anttila et al., 2022), then manually validate these in terms of their semantic, predictive, and convergent validity (Quinn et al., 2010; Grimmer and Stewart, 2013). For our topic models, we use latent Dirichlet allocation (LDA; Blei et al., 2003), a hierarchical admixture model of text, which allows us to capture and convey the prevalence of various topics, so that we can contrast these across news sources, and over time. We perform topic modeling with MALLET (McCallum, 2002), and to pre-process the articles, we lemmatize all tokens with WordNet's morphy feature (Miller, 1995), and also extract common bigrams. The resulting unigrams and bigrams were then converted to term–document matrices and provided as inputs to MALLET. We created models exploring varying numbers of automatically discovered topics, we subsequently evaluated the coherence score (Mimno et al., 2011) of the resulting topics and manually spot-checked them. We estimate a topic's prevalence in a news source or year by averaging the topic's weight across the articles from that source or year. For the period 1998–2020, we operationalized issue framing by creating models, setting the number of topics from k = 5 to 20, and evaluating the coherence of the resultant topics. We found that using six topics produced the highest coherence score, and we identified each of these topics with an identifying label.

[1] See Appendix C.

3.2 Sentiment Analysis

Sentiment score may provide evidence of stylometric differences between newspapers sources, so we apply computational sentiment analysis to measures the tone and connotations of articles. While it is common to use hand-crafted sentiment (valency) lexica (Mohammad, 2018), we selected a technique that is robust to the specific words that are chosen. We chose a BERT-based model to classify a given sentence as positive or negative because of its near state-of-the-art sentiment classification abilities. We treat sentiment as a binary attribute (+, −) and use a probabilistic classifier trained on the Stanford Sentiment Treebank (SST-2; Socher et al., 2013).[2] The model uses DistilBERT (Sanh et al., 2019) for feature extraction from text. Together with the analysis of lexical usage and topic modeling, sentiment analysis strengthens the understanding of newspapers' portrayal of the Hong Kong protests.

3.3 Lexical Frequency

Word frequency exposes discrepancies in word choice and usage. A lack of event-related keyword in contemporaneous articles from different newspapers may signal omission of events in some of them. Analysis of variance (ANOVA) is a class of sampling theory–based methods for comparing the means of a quantitative response variable, when the explanatory variable is categorical (Agresti, 2017). A statistically significant p-value supports that the means of both populations are different. As our corpus dis- plays a non-parametric distribution, we apply Mann– Whitney U , splitting by newspaper source and using the Holm–Bonferroni correction with significance level of $\alpha = 0.01$, to test whether any of the 19 protests-related keywords has statistically significant differences in usage (i.e., *confront*, *confrontation*, *crackdown*, *democracy*, *freedom*, *freedom of speech*, *independence*, *occupation*, *protest*, *protests*, *resistance*, *rights*, *riot*, *rule of law*, *severe*, *tension*, *terrorism*, *terrorist*, *unrest*).

The Mann–Whitney U, splitting by newspaper source and using the Holm–Bonferroni correction, shows that every word has statistically significant differences in usage except severe. The same test, splitting by before and after June 2019, shows statistically significant differences only for five (out of the 19) keywords: *protests*, *unrest*, *rights*, *rule of law*, and *democracy*. For the Friedman's test with four categories – that is "west before June"; "west after June"; "Hong Kong before June"; and "Hong Kong after June" – no keyword showed statistically significant differences.

[2] There is merit to including a third 'it's complicated' class (Kenyon-Dean et al., 2018).

3.4 Collocations

Collocations further the understanding of how words are used differently across news sources. A word embedding model is a smoothed model of collocation that employs a vectorial, rather than symbolic, representation of words. It seeks to assign similar vectors to words in similar contexts, and different vectors to words in different contexts. If the usage of a word changes, then this should be reflected in changes to the word's context and thus changes in the word's embedding (Kulkarni et al., 2015). To assess these semantic and collocational shifts, we both replicate and extend the difference-in-usage model of Gonen et al. (2020).

1. Partition the corpus C into C_a and C_a based on the attribute of interest a.
2. Fit separate word embedding models for each partition: M_a and M_a.
3. Select a keyword w of interest.
4. Obtain the set of nearest neighbors $NN_a(w)$ and $NN_a(w)$ of w according to M_a and M_a.[3]
5. Score the usage-change of w as the size of the intersection, $|NN_a(w) \cap NN_a(w)|$.

After this process, if a word w is used differently based on the presence or absence of the attribute, we expect its score to be low. Words whose usage does not depend on a will have similar neighborhoods in each split. To extend Gonen et al. (2020), we contextualize the similarity score of a given word with the percentile in which the score falls. This distributional measure is more interpretable than the raw similarity score.

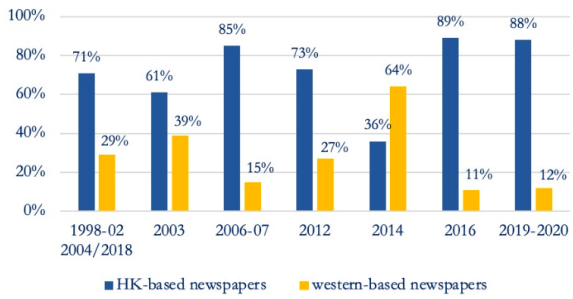
[3] Following Wendlandt et al. (2018) and Gonen et al. (2020), we use 1000 nearest neighbors.

4 Discussion

4.1 Who Covers the Protests in Hong Kong?

The number of articles about protests in Hong Kong allows us to gauge the news value of protests as well as the thematic relevance to newspapers in general. Between 1998 and 2020, Hong Kong–based newspapers published more articles about protests in Hong Kong than western-based newspapers, except for 2014 when this trend was reversed.

Figure 1. Hong Kong– and western-based newspapers' coverage of selected episodes of civic unrest between 1998 and 2020



Western-based newspapers' coverage of protests in Hong Kong is punctuated by sharps peaks and dips, and declines over time, most significantly between 2014 and 2019–2020. With 425 articles, the NYT published more than half than what all western-based newspapers published on Hong Kong protests over the 22-year time- line of our research. With 3347 articles, the SCMP published almost 8.8 times more articles than the China Daily (i.e., 3347 vs 381), and more than any other news- papers in our sample. Moreover, the SCMP and China Daily together published 4.7 times more articles than the NYT, WSJ, WaPo, FT, the Guardian, the Times of London combined (i.e., 3728 vs 793). As the main English-language outlets in Hong Kong, it is not surprising that the SCMP and China Daily coverage of Hong Kong matters is more frequent and in-depth than that of western-based newspapers, particularly nowadays as newspapers have reduced the space, resources and commitment devoted to a range of topics, and have especially cut back on foreign news.

4.2. What is the Tenor of Articles about the Protests?

We corroborate the findings of more negative tone in Hong Kong–based newspapers between 1998 and 2020. At 36.9% the Hong Kong–based articles' average positivity is slightly lower than the 38.1% of western-based articles. There is, though, wide variation across sources. At 31.3% and 31.5% the SCMP and the Times emerge as the newspapers with the most negative tone over- all, even though the SCMP published the most articles, whereas the Times rarely publishes about the protests. Both The Guardian,

at 32.9%, and the Financial Times, at 33.9%, also rank low in positivity, which makes UK- based newspapers the more negative about Hong Kong protests among all western-based newspapers. US-based newspapers average a positivity score of 36%, with the NYT articles being almost imperceptibly more positive than the WSJ (i.e., 36.3% vs 36.1%) and WaPo (i.e., 36.3% vs 36%). At 40%, the China Daily has the most positive tone among both Hong-Kong– and western-based newspapers. Finally, Hong Kong–based newspapers articles' average positivity remains lower than that of western-based newspapers articles in 2014 (i.e., 33.2% vs 35.7%), and also in 2019–2019 (i.e., 31.4% vs 32.9%).

4.3. What Do Headlines Hint about the Protests?

News headlines are bait. They are meant to catch readers' attention by using narrative mechanisms and sensational or provoking words (Blom and Hansen, 2015), and help the reader get the most out of the news with minimum effort (Dor, 2003). We tested for the presence of long, short, and judgmental headlines vis-a`-vis protests in Hong Kong. Sixty-three percent (63%) of articles in the corpus have long headlines (i.e., include six or more words), whereas the remaining 36% have headlines with less than six words, with these trends not varying significant over the extended timeline of the research. With headlines like "The Worst of Times" or "Hong Kong: A City Divided" the NYT emerges as the newspaper with the highest likelihood of having telegraphic headlines (i.e., 6.3 times more likely), whereas with headlines like "Hong Kong Extradition Bill: Business Groups Breathe Collective Sigh of Relief Over Government Decision to Delay Legislation" the SCMP is the least likely of the newspapers to have short headlines (i.e., 11% less likely).[4] The NYT emerges as the one newspaper whose headlines offer a clear and dramatic view of what the article is about it to stimulate its readership's curiosity. SCMP long headlines showcase key information from the articles, and in consistently doing this, the SCMP emerges as the newspapers that more efficiently succeeds at both story summarization, immediacy satisfaction, and attention direction among all newspapers.

Headlines can be structurally classified as either verbal or nonverbal. Some 75% of the headlines in the NYT articles were nonverbal, while only 25% of them were verbal; most of the nonverbal headlines were modified— i.e., they may include a term that adds descriptive in- formation to the headline (Quirk et al., 2010, 65). Furthermore, on average, about 53% of the headlines of western-based newspapers other than the NYT were also of the nonverbal kind. We also found that, in their headlines, western-based

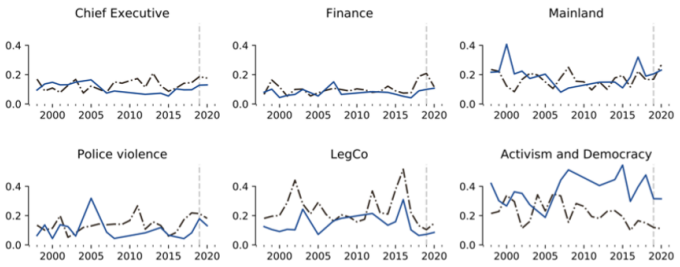
[4] The full regression model containing all predictors was statistically significant, X2 (8; N = 4522) = 723.787, p ¡ 0.001. The model correctly classifies 75% of cases, and explains between 16.9% (Cox and Snell R2) and 21.5% (Nagelkerke R2) of the cases. The strongest predictor for short headlines is the variable for the NYT, with an Exp(β) of 6.3, whereas the weakest predictor for short headlines is the variable for the SCMP, with an Exp(β) of − 0.89.

newspapers used "presupposition" (van Dijk, 1995, 273) (Bonyadi and Samuel, 2013, 5) 45% more than Hong Kong–based newspapers to posit a negative attribute for what articles identify as others (e.g., Hong Kong Chief Executive; Hong Kong government; Beijing; China; police) and positive ones for us (e.g., protesters; citizens; rights; freedoms).

4.4. How Do Newspapers Frame the Protests?

Our unsupervised topic modelling reveals that, both over time and in the case of specific protest events (i.e., the anti-ELAB protests) Hong Kong– and western-based newspapers use the same topics. The prominence and timing of how the same topics are used is, however, different. As such, what emerges from the topic modelling analysis should be understood as journalistic frames, unique and specific to western- and Hong Kong–based newspapers' coverage of the Hong Kong protests between 1998 and 2020.[5] The treatment of the police violence topic/frame helps showcase the role and relevance that factors such as norms and practices of the news industry, newspapers' desire to appeal to their own readership, preference for big picture issues, and/or focus on the details of domestic issues play in shaping the narrative of protests coverage.

Figure 2. Principal topics/frames used in protest news construction in Hong Kong– and western-based newspapers, 1998–2020. Solid blue: Western. Dashed black: HK.



Topic	Top 10 words
Chief executive	bill, lam, extradition, public, court, executive, legal, case, cheng
Finance	cent, per_cent, hk, business, company, market, million, property, trade, billion
Mainland	beijing, chinese, country, system, state, mainland, national, law, foreign, central
Police violence	officer, station, violence, force, arrested, yesterday, students, road, university, day
Legislative council (LegCo)	election, party, leung, council, candidate, lawmaker, vote, executive, camp, legislative
Activism and democracy	student, n't, movement, street, Chinese, leader, mr, day, beijing, democracy

Police violence. Between 1998 and 2018, the trends for how the police violence frame has been used in western- and Hong Kong–

[5] TADA 2021, 11th Annual Conference on *New Directions in Analyzing Text as Data. Panel Longitudinal Studies of Language*; discussant Philip Resnik. https://tada2021.org

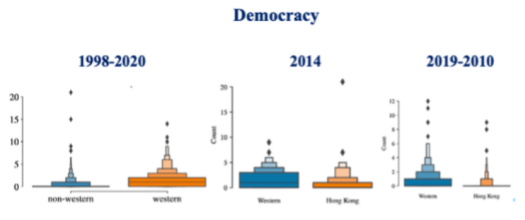
based newspapers hardly mimic each other or move in opposite directions, as between 2003 and 2007, when the use of this frame peaks in western-based newspapers and dips for Hong Kong–based ones, or 2008 and 2012, when the opposite happens. The trends, though, mimic each other in Hong Kong– and western-based newspapers between 2019–2020. This points to the police violence frame having equal significance for both sets of newspapers when discussing the protest-related cycle of violence that the Hong Kong government could not break.

4.5. How Often Do Newspapers Use Protest-Relevant Terms?

The investigation of the evolution of how words are used differently, both in the western/Hong Kong split and over time, reveals that, with the exception of 2019– 2020, western-based newspapers have used the terms **democracy** and **freedom** more often than Hong Kong– based newspapers.

Between 1998 and 2020, freedom appears 492 in western-based newspapers and 70 times in Hong Kong– based newspapers. In the same time period, democracy appears 242 times in western-based newspapers and 107 in Hong Kong–based newspapers. In 2014, freedom appears 127 times in western-based and only 18 times in Hong Kong–based newspapers, whereas democracy appears 34 times more in western-based than Hong Kong–based newspapers (i.e., 416: 12). These trends are reversed between 2019 and 2020, when freedom appears 755 times in Hong Kong–based newspapers and 365 in western-based ones. Democracy appears 668 times in Hong Kong–based newspapers and 217 times in western-based ones.

Figure 3. Lexical frequencies for democracy



	Western-based newspapers	Hong Kong-based newspapers
1998-2020	875/793 = 1.103	787/ 3729 = 0.211
2014	416/236 = 1.762	12/131 = 0.091
2019-2020	217/390 = 0.556	668/3009 = 0.222

However, as shown in Figure 3, the frequencies of use of democracy and freedom are, overall, lower in Hong Kong than in western-based newspapers. Moreover, western-based newspapers use democracy and freedom predominantly as a noun, whereas Hong Kong–based newspapers tend to use both terms more often as qualifiers rather than as nouns.

The difference in frequencies may be partially rooted in the type of articles that the newspapers publish. Western- based newspapers tended to cast citizens' civic assertiveness as their fight for democracy and the freedoms that come with it, or resistance against authoritarian tightening that Hong Kong has been experiencing following the 1997 handover. This narrative may require a more frequent

use/discussion of democracy and freedom as concepts and values. On the other hand, Hong Kong–based newspapers tend to focus their discourse narrowly on the details of the protests rather than on their meaning. With such specific narrative, it is, perhaps, not surprising that democracy and freedom are used sparingly and as qualifiers across the large number of articles published.

4.6. How are terms used differently between the West and Hong Kong?

The analysis of lexical usage reveals semantic divergence in certain keywords between Western- and Hong Kong–based newspapers between 1998 and 2020, and also with regards to particular episodes of protests. As shown in Table 3, between 1998 and 2020, the most significant semantic divergence is found in the lexical usage of the words ***riot*** (98th percentile), ***protest*** (88th percentile), ***occupation*** (80th percentile), ***confrontation*** (70th percentile), ***tensions*** (59th percentile), and ***crackdown*** (51st percentile). Moreover, a visual inspection of the term's nearest neighbors for the western-based model suggests the prevalence of neutral or descriptive lexicon as in the case of *scene, clearance, dispersal, crowds* for the word **riot**; *sit-ins, rally, campaign* for the word **protest**; or *dispute, turmoil, uncertainty* for the word **tensions**.

Table 1. Neighbors in 1998–2020 for select protest- related keywords

Hong Kong–based	Western-based
riot (98th percentile)	
fired, spray, barricades, officers, pepper, station, rubber, cocktails, firing, teargas	mob, rampage, mobs, siege, scene, clearance, dispersal, crowds, clash, radicals
protest (88th percentile)	
activists, peaceful, rally, mass, organizers, demonstrations, streets, occupied, admiralty, main	demonstration, sit-ins, rally, demonstrations, campaign, rallies, march, movement, sit-in, protests
occupation (80rd percentile)	
denounce, supporters, peacefully, confrontation, join, occupying, radical, momentum, peaceful, anti-government	mayhem, rallies, demonstrations, demonstration, marches, sit-ins, outbursts, bloody, 79-day, scenes
confrontation (70rd percentile)	
peacefully, break, dramatically, preparing, standoff, storm, driving, demonstrate, chaotic, dislodge	turning, stand-offs, mayhem, resorting, confrontations, extreme, resorted, quickly, disruptive, chaotic

In contrast, the nearest neighbors in the Hong Kong–based model relate to adversarial or hostile behaviour as in the case of fired, barricades, pepper, teargas for the word riot; break, standoff, storm, chaotic, dislodge for the word confrontation. These trends are evidence of Hong Kong–based newspapers' choice of the protest paradigm when publishing about civic unrest in Hong Kong, and also that the SCMP and China Daily reporting about protests has remained the same, although Hong Kong protests have evolved over time. Moreover, the fact that the most significant semantic divergence is found in in the lexical usage of nouns used in their singular form, suggests that Hong Kong–based newspapers' consistent use of the protest paradigm, to

frame the discussion of protests in Hong Kong, could be a strategy used to criticize the values embodied in those nouns while reporting about them more distantly as objects or actions.

Table 2. Neighbors in 2014 for select protest-related keywords

Hong Kong–based	Western-based
occupation (75th percentile)	
join, started, even, protesting, threat, thought, umbrella, planning, probably, revolution	a, court, support, work, admiralty, go, even, legal, protest, they
protest (61st percentile)	
movement, main, pro-democracy, student, peace, group, district, site, admiralty, love	court, admiralty, even, a, pan-democrats, occupation, social, three, ?, way
confrontation (53rd percentile)	
scene, losing, showing, despite, grew, avoid, businesses, families, workers, demonstration	line, lai, came, much, wong, democratic, number, democracy, sit-in, still
tensions (24th percentile)	
became, questions, laws, prevent, little, half, winning, helped, closely, internal	participants, views, rights, meant, yesterday, month, also, city, protesters, students

Tables 1 and 2 show that also in the case of 2014 the Occupy Central protests and the 2019–2020 anti-ELAB protests the analysis of lexical usage reveals semantic divergence in certain keywords between western- and Hong Kong–based newspapers that are consistent with what found for the 1998–2020 period. In the case of the 2014 Occupy Central protests, the most significant semantic divergence is found in the lexical usage of ***occupation*** (75th percentile), ***protest*** (61st percentile), ***confrontation*** (53rd percentile), whereas the least significant semantic divergence is found for some of the very words that are used most differently over time (i.e., tensions (24th percentile); riots (20th percentile); and crackdown (18th percentile).

Table 3. Neighbors in 2019–2020 for select protest- related keywords

Hong Kong–based	Western-based
confront (17th percentile)	
retreat, intimidated, abused, reminded, understandable, upset, confronting, provoked, regularly, provoke	work, met, acts, spirit, reasons, decisions, trying, voice, intolerable, tsang
protest (66th percentile)	
rally, sit-ins, demonstration, demonstrators, rallies, campaign, strike, movement, march, demonstrations	demonstrations, umbrella, peaceful, movement, began, streets, activists, mass, 1m, referendum
tensions (63rd percentile)	
us-china, tension, war, dispute, uncertainty, heightened, prolonged, worsening, fallout, turmoil	culture, state-owned, protections, tourists, market, base, rise, travel, closer, argued

As for the 2019–2020 anti-ELAB protests, the most significant semantic divergence is found in the lexical usage of ***protests*** (66th percentile) and ***tensions*** (63rd percentile) (Table 3). On the one hand, the consistency of the prevalence of neutral or descriptive lexicon for the Western-based models, and the recurrence of adversarial or hostile behavior lexicon for the Hong Kong–based model, and the fluctuations in the magnitude of the semantic divergence in certain keywords when comparing across protests, are likely to be linked to the characteristics specific of the various episodes of

protests. On the other, they may be explained by Hong Kong media "norms of political correctness" (Lau and To, 2002, 74) vis-a-vis Beijing, or the "strategic rituals" (Lee, 2000, 317) Hong Kong newspapers have established to counter Beijing's "strategic ambiguity" (Cheung, 2003) and ensuing self-censorship, or by cultural co-orientation, resulting from Hong Kong journalists' views shifting closer to China's official views.

4.7. Does Coverage Differ Before and After the Onset of Protests in June 2019?

We investigated whether there are differences in these differences over time, in the 2019–2020 anti-ELAB protests. We found that, over time, the semantic context of the protest keywords becomes more polarizing and intense. Statistical analysis lets us compare the means of a continuous response variable, modulated by two categorical explanatory variables. We use the Holm–Bonferroni correction to mitigate false discovery. In our case, the explanatory variables are the source (western/HK-based newspapers) and the date: was the article published before or after July 1, 2019?

In the case of unrest, democracy, rights, crackdown, and protest our analysis found significant differences in the way these terms were used in newspapers before and after July 1, 2019. For the Friedman's test with four categories (western-based newspapers, Hong Kong– based newspapers) x (before, after) no keyword showed significant differences. These results suggest that any already existing biases were not discernibly altered by the onset of the anti-ELAB protests. Moreover, building on the richness of our dataset, we also sought to quantify the degree to which the intro- duction of ELAB acted as a pivotal moment in how newspapers portray the Hong Kong protests, and found that June 2019 emerges as a turning point, after which the meaning of several keywords shifts for at least the remainder of 2019.

We split the corpus into "pre-June 30th, 2019" and "post-June 30th, 2019" to investigate whether the way in which Hong Kong– and western-based newspapers portrayed episodes of civic unrest differently following the protests and demonstrations that took place over the month of June 2019. Neighborhood shift analysis revealed significant low scores for resistance, severe, riots, confront, confrontation, and terrorism, which suggests that the context and/or semantic meaning for these words changes from early to late 2019, regardless of whether Hong Kong– or western-based news sources are considered. For instance, in the first half of 2019, neigh- bours for riots include terms like actions, open, engage, and taken, which that are not charged, and in the context of either a reporting or an opinion piece descriptively informs readers. However, in the second half of 2019, the nature of neighboring terms for riots changes to include more polarizing terms such as violent, escalated, destructive, triggered, anti-government, and sparked. Similarly, pre-July 2019, neighbors for terrorism include, among others, terms

like covered, lawyers, and negative. Post-June 2019, neighboring words become politically charged, and include criminals, destructive, extreme, lawless, punishing, and barbaric.

These findings reflect well the extent to which June 2019 was a pivotal moment in the context of the 2019–2020 Hong Kong protests. As protests escalated exponentially during the month of June, feelings of social danger pre- vailed in newspapers' accounts of the events. Citizens' civic assertiveness was described more and more harshly over time. Our dataset allowed to see that articles pre- July 2019 focused on general descriptions of protesters' tactics, whereas post-June 2019 on detailed description of violent actions that took place during the protests as well as mentions of the negative social impacts that such actions may cause.

Table 4. Neighboring terms for the word riot in Hong Kong–based newspapers in 1998–2020, 2014, and 2019– 2020

1998–2020	2014	2019–2020
98th percentile	20th percentile	99th percentile
fired, spray, barri-cades, officers, pep-per, station, rubber, cocktails, firing, tear-gas	batons, fired, shield, canisters, umbrellas, rubber, disperse, bul-lets, officers	violent, escalated, destructive, triggered, anti-government, sparked

However, 2019 was not the first time that the semantic context of protest keywords had become more polarizing and intense. As Table 6 shows, between 1998 and 2020 as well as in 2014, the nature of neighboring terms for the word riot was the same as that found in Hong Kong– based newspapers in the second half of June 2019. The inefficacy of the Hong Kong government's response to the 2019–2020 protests, and the Hong Kong's shrinking freedom of speech that may help explain why, in 2019, highly polarized protest keywords impacted Hong Kong in such consequential way, whereas their impact was barely noticed in 2014, or between 1998 and 2020.

5. Conclusions

We show how powerful the curated dataset of 4522 articles, spanning over a 22-year time horizon of the Hong Kong Protest News Dataset can be in revealing longitudinal and synchronic changes in how Hong Kong– and western-based news sources cover protests in Hong Kong.

The sheer volume of the articles in our dataset validates the news value of Hong Kong protests for both Hong Kong– and western-based newspapers, and shows that coverage of Hong Kong–based newspapers remains consistent and sustained over time, whereas that of western- based newspapers is punctuated by sharps peaks and dips, declining between 2015 and 2020. We speculate that these differences create an opportunity for the Hong Kong–based press to set the agenda for how protests are

framed and reported. Within these results, 2014 emerges as an outlier, with western-based newspapers publishing twice as many articles as Hong Kong–based ones on Occupy Central protests.

We prove diachronic consistency between the topics/frames that western- and Hong Kong–based news- papers use to cover the protests in Hong Kong, which points to the generalizability of our findings. We also found that newspapers rely on a limited set of frames, which portray protests as deviant actions characterized by violence and vandalism and detrimental for society. The analysis of stylometric differences across newspapers sources shows evidence of a more negative tone in Hong Kong–based newspapers when reporting about the protests in Hong Kong, and that western-based news- papers use the terms democracy and freedom more and predominantly as nouns, whereas Hong Kong–based newspapers use them less frequently and, generally, as qualifiers rather than as nouns.

Our investigation of word embedding neighborhoods broadened the current understanding of how words are used differently between Western- and Hong Kong– based newspapers. We confirmed semantic divergence in certain keywords both between Western- and Hong Kong–based newspapers over time and with vis-a`-vis particular episodes of protests. We confirmed that, over time, the semantic context of the protest keywords be- came more polarizing, and that June 2019 is as a pivotal moment, after which the meaning of several keywords shifts for at least the remainder of 2019.

Finally, the extended time horizon of the Hong Kong Protest News Dataset also allowed us to capture how the semantics of protest keywords became more polarizing and intense during episodes of protest beyond the 2019– 2020 anti-ELAB ones. We show that between 1998 and 2020 as well as in 2014, the neighboring terms for the word riot were remarkably similar to those Hong Kong- based newspapers used in 2019. We hypothesize that the differing impact of similarly polarized protest keywords over time can be explained by "shifts in journalistic paradigms" (Chan and Lee, 1984, 97) which altered the boundaries of press freedom in Hong Kong.

References

Agnone, J. (2007). Amplifying Public Opinion: The Policy Impact of the U.S. Environmental Movement. Social Forces, 85(4):1593–1620, 06.

Agresti, A. (2017). Statistical methods for the social sciences. Pearson.

Baden, C., Pipal, C., Schoonvelde, M., and van der Velden, M. A. C. G. (2021). Three gaps in computational text analysis methods for social sciences: A research agenda. Communication Methods and Measures, 0(0):1–18.

Bhatia, A. (2015). Construction of discursive illusions in the 'umbrella movement'. Discourse & Society, 26(4):407–427.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. J. Mach. Learn. Res., 3(null):993–1022, March.

Bonyadi, A. and Samuel, M. (2013). Headlines in news- paper editorials: A contrastive study. SAGE Open, 3(2):2158244013494863.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). Good News Everyone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1554– 1566, Marseille, France, May. European Language Resources Association.

Boydstun, A. E., Card, D., Gross, J., Resnick, P., and Smith, N. A. (2014). Tracking the development of media frames within and across policy issues.

Boykoff, J. (2006). Framing dissent: Mass-media coverage of the global justice movement. New Political Science, 28(2):201–228.

Bu¨yu¨ko¨z,B., Hu¨rriyetog˘lu,A.,andO¨zgu¨r,A.(2020). Analyzing ELMo and DistilBERT on socio-political news classification. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pages 9–18, Marseille, France, May. European Language Resources Association (ELRA).

Chan, J. M. and Lee, C.-C. (1984). The journalistic paradigm on civil protests: A case study of Hong Kong. The news media in national and international conflict, pages 183–202.

Chan, C. K. (2015). Contested news values and media performance during the umbrella movement. Chinese Journal of Communication, 8(4):420–428.

Cheung, A. S. (2003). Hong Kong press coverage of China–Taiwan cross-straits tension: Anne Sy Cheung. In Hong Kong in transition, pages 219–234. Rout- ledge.

Du, Y., Zhu, L., and Yang, F. (2018). A movement of varying faces: How "occupy central" was framed in the news in Hong Kong, Taiwan, mainland China, the UK, and the U.S. International Journal of Communication, 12(0).

Earl, J., Martin, A., McCarthy, J. D., and Soule, S. A. (2004). The use of newspaper data in the study of collective action. Annual Review of Sociology, 30(1):65–80.

Federico, M., Giordani, D., and Coletti, P. (2000). Development and evaluation of an Italian broadcast news corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece, May. European Language Resources Association (ELRA).

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3570–3580, Brussels, Belgium, October-November. Association for Computational Linguistics.

Gamson, W. A. and Wolfsfeld, G. (1993). Movements and media as interacting systems. The ANNALS of the American Academy of Political and Social Science, 528(1):114–125.

Gonen, H., Jawahar, G., Seddah, D., and Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 538–555, Online, July. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

King, B. G. (2014). The Tactical Disruptiveness of Social Movements: Sources of Market and Mediated Disruption in Corporate Boycotts. Social Problems, 58(4):491–517, 07.

Lau, T.-y. and To, Y.-m. (2002). Walking a tight rope: Hong Kong's media facing political and economic challenges since sovereignty transfer. In Ming K. Chan et al., editors, Crisis and Transformation in China's Hong Kong, page 322. Hong Kong University Press.

Lee, C.-c. (2000). The paradox of political economy: Media structure, press freedom, and regime change in Hong Kong. Power, money, and media, pages 288– 336.

Lee, F. L. F. (2014). Triggering the protest paradigm: Examining factors affecting news coverage of protests. International Journal of Communication, 8(0).

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass. edu.

McCarthy, A. D., Scharf, J., and Dore, G. M. D. (2021). A mixed-methods analysis of western and Hong Kong–based reporting on the 2019–2020 protests.

In Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 178–188, Punta Cana, Dominican Republic (online), November. Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A lexical database for English. Commun. ACM, 38(11):39–41, November.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 262–272, Edinburgh, Scot- land, UK., July. Association for Computational Linguistics.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 174–184, Melbourne, Australia, July. Association for Computational Linguistics.

Papanikolaou, K. and Papageorgiou, H. (2020). Protest event analysis: A longitudinal analysis for Greece. In Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020, pages 57–62, Marseille, France, May. European Language Resources Association (ELRA).

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Black- burn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Quirk, R., Greenbaum, S., Leech, G., and Svatvik, J. (2010). A comprehensive grammar of the English language. Longman, London.

Rawnsley, G. D. and Rawnsley, M.-Y. T. (2002). Political communications in greater China. Curzon.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. Commun. ACM, 18(11):613–620, nov.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

Scharf, J., McCarthy, A. D., and Dore, G. M. D. (2021). Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In Proceedings of the 4th Work- shop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), pages 43–52, Online, August. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Solomon, W. S. (1996). Covering Dissent: The Media and the Anti-Vietnam War Movement. By Melvin Small. (New Brunswick: Rutgers University Press, 1994. x, 228 pp. Cloth, 42.00,ISBN0–8135–2106–8.Paper, 16.00, ISBN 0-8135-2107-6.). Journal of American History, 82(4):1651–1651, 03.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology, 29(1):24–54.

Tsfati, Y. and Walter, N. (2019). The world of news and politics. Media Effects: Advances in Theory and Research.

van Dijk, T. A. (1995). Discourse semantics and ideology. Discourse & Society, 6(2):243–289.

Weiss, M. L. and Aspinall, E. (2012). Student activism in Asia: Between protest and powerlessness. U of Minnesota Press.

Wendlandt, L., Kummerfeld, J. K., and Mihalcea, R. (2018). Factors influencing the surprising instability of word embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the As- sociation for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2092–2102, New Orleans, Louisiana, June. Association for Computational Linguistics.

Wong, H. T. and Liu, S.-D. (2018). Cultural activism during the Hong Kong umbrella movement. Journal of Creative Communications, 13(2):157–165.

Wong, M. Y. (2021). Democratization as institutional change: Hong Kong 1992–2015. Asian Journal of Comparative Politics, 6(1):92–106.

Yu, M. (2015). Framing Occupy Central: A content analysis of Hong Kong, American and British newspaper coverage.

Appendixes

A Primer on Mixed-method Research

Mixed-method research paradigms (MMR) are a class of research where quantitative and qualitative research techniques are combined into a single study. Philosophically, the MMR paradigm uses a logic of inquiry that includes the use of induction, deduction, and abduction, which allows the researcher to approach the research questions in a way that offers the best chance of obtaining useful answers (Chatterjee, 2013). It is acknowledged that both quantitative and qualitative analyses suffer from certain specific shortcomings, which are not so much intrinsic to the two designs, but result from the lack of resources many researchers face, and forces them to prioritize either scope or depth. A mixed-methods design aims to combine the advantages of both methods in one single framework. Hence, the different strong points of the methods are expected to overshadow or even push away the weaknesses.

The attention to MMR is relatively new and can be traced back to Achen and Snidal (1989)'s recommendation to use historical case studies as a useful complement to statistical research; their plea was further strengthened by Verba's work in the early 1990s (Verba et al., 1993, 1995; Verba, 1996), and particularly by Tarrow (1995), which openly called for bridging qualitative and quantitative modes of research in social science. In the last decade, a great number of authors have made a 'quantum leap' (Levy, 2007) in social science methodology by providing a highly structured post-positivist approach to qualitative analysis. In so doing, they have enriched social science with important methodological innovations (Coppedge, 1999; Gerring, 2004; Lieberman, 2010).

Small-N samples allow to arrive at outcomes that are qualitatively thick and empirically well-grounded, possess a high order of complexity, and therefore relevant in bounded times and places; in addition, when used over an extended time horizon they strengthen the methodological validity of the research exercise, and of its results (Tipton et al., 2016; Smith and Little, 2018; Konietschke et al.,2021).

B Media Landscape

B.1 The Hong Kong Media Landscape

Hong Kong has a long legacy of an aggressive and boisterous media, in both the dominant Cantonese language and English. Newspapers are widely read, and they often carry sharp critiques of government and police failures. The forces of de-colonization and globalization have shaped the development of the

English-language news media in Hong Kong over the past two decades. When in the mid-1980s Hong Kong began the transition to Chinese sovereignty, the position of influence of the local English-language press began to erode while the Chinese-language newspapers started to proliferate and expand their readership. As Hong Kong moved out of the colonial era, however, a counter-trend emerged with international English- language media conglomerates, mostly US-owned. The English-language news media in Hong Kong is very influential, and market mechanisms do not explain why this minority of readers could wield such disproportionate influence, even if it less than it used to be. This skewed influence is largely due to the fact that English was the only official language of Hong Kong until 1974 and remains one of the most used languages following the 1997 handover, and the most important means for upward mobility. In terms of plurality of media voices, the Hong Kong media landscape fits with the findings of Djankov et al. (2003), which show that almost universally the largest media firms are owned by the government or by private families. Government ownership is more pervasive in broadcasting than in the printed media. The landscape also fits with 1Gehlbach and Sonin (2014) who explain government control of the media and variation in media freedom across countries and over time.

The South China Morning Post was founded in 1903 by an Australia-born Chinese and a British journalist with funding from mostly non-Chinese businesspeople, with the manifest goal of supporting the reform movement in China, where revolutionaries sought to overthrow the imperial 1037 Qing dynasty. The newspaper has long been the broadsheet of the city's elites, and it is arguably the city's most important title internationally, a position gained from a combination of both its size and its ownership. Through the late 1980s and into the mid-2000s, the paper was owned by the media tycoon Rupert Murdoch and then the Malaysian billionaire Robert Kuok. In 2015, the SMCP was acquired by Alibaba with the "uncertain future for traditional publishing" as a key reason behind the sale. As Hong Kong's leading English-language newspaper, the South China Morning Post reports on issues and topics that are considered sensitive in mainland China, where the websites of several international media are blocked. While Ma is known to be politically well-connected, the shift in ownership is not as drastic as some people have indicated. Previous owners, in fact, were business tycoon with close ties to the Chinese government. The SCMP is not as well read as the international outlets that it would like to compete with, but because of its unique position—as the main English-language outlet in a strategically important city—its coverage plays an outsize role in shaping international understanding of events not just in Hong Kong but across the border in China, as well. Moreover, its coverage is far more credible than any mainland outlet, and has been courting a global readership hungry for news from China by dropping its paywall. In 2018, it announced a tie-up with Politico signaling the newspaper's "growing credibility and authority." (Gary Liu, SCMP's CEO, internal email). Thanks to those factors, as well as drastically increased interest in China, where, of course, the coronavirus pandemic began, the SCMP has seen a sharp rise in readership.

Though its daily print circulation is relatively limited, at just over 100,000, it averages more than 50 million monthly active users—a tenfold increase over the past three years—and nearly 200 million pageviews a month.

The China Daily is part of the China Daily Group, which runs 16 print publications in China and abroad, and according to its mission statement, 'is an authoritative provider of information, analysis, comment and entertainment for global readers with a special focus on China.' The China Daily is the largest English-language daily in China, and although it is state-owned, it is not officially a mouthpiece of the ruling Communist Party and is considered more liberal than the other Chinese state newspapers that circulate in Hong Kong. The China Daily Group also publishes the Hong Kong, US, European, African, Asian and Latin American editions of China Daily, with, according to its own statistics, a total circulation of 900,000 copies. China Daily in mainland China is published on a daily basis; there are two different forms of appearance outside the mainland. There is a China Daily Hong Kong edition and a China Daily USA edition, both with daily frequency, and there are weekly editions for Asia, Europe, North America, Africa and Latin America. The newspaper also publishes China Watch, which is circulated as a supplement with the Washington Post, Los Angeles Times and London's Daily Telegraph.

B.2 The North American and British Media Landscape

North American and British media landscapes have been characterized by the centrality of large-scale cultural industries since the development of the penny press in the 1830s. For several decades in the mid-twentieth century, an equilibrium existed in the media system, with stable markets that made the dominant media companies highly profitable and very influential as social institutions. Newspapers invested heavily in newsrooms, and the journalist profession grew in autonomy and influence. Journalism was characterized by a low level of political parallelism, with the norm of objectivity dominating journalistic ethics, and most news organizations avoiding identification with political parties. Over time, economic, technological, and political change began disrupting those very elements that guaranteed the stability and fairness of the press system in the late twentieth century. Stable boundaries that once separated markets have been disrupted by digital convergence and deregulation. Changing business models have impacted newsrooms' sizes, and the Internet has accelerated the fragmentation of broadcast industries. Political parallelism has increased, and newspapers have become more inclined to adopt partisan identities, with little concern about the reporting pitfalls of adopting such clear-cut political orientation, including declining public trust.