The following publication Tay, D. (2021). Automated lexical and time series modelling for critical discourse research: A case study of Hong Kong protest editorials. Lingua, 255, 103056 is available at https://dx.doi.org/10.1016/j.lingua.2021.103056.

Automated lexical and time series modeling for critical discourse research: a case study of Hong Kong protest editorials

Abstract

This paper advances a novel approach to critical synchronic and diachronic discourse analysis using automated lexical and time series modeling. It is illustrated by a case study of near-daily editorials (N=201; 300,081 words) from 9 June to 2 October 2019 on the Hong Kong protest movement in three ideologically contrasting sources – *China Daily* (CD), *South China Morning Post* (SCMP), and *Hong Kong Free Press* (HKFP). Lexical analysis with Linguistic Inquiry and Word Count (LIWC) first revealed four predominant socio-psychological word categories - relativity, drive, cognitive, and affect. Overall, HKFP expresses anger at the government, CD lays blame on protestors' violent actions, and SCMP occupies a middle position to focus on less political aspects. Time series modeling is then applied to redirect attention from these aggregated differences to how they unfold day-to-day. It was found that while positive affect words are characterized by short-term consistencies and fluctuations, most variables exhibit random variation across time. The approach allows precise description of how linguistic variables in neighboring time periods inter-relate, offering rich interpretative possibilities for different linguistic/discourse contexts. Furthermore, determining whether a variable is 'modelable' offers a systematic and replicable way to interrogate the assumption that discourse inevitably serves to construe social reality.

Keywords: LIWC, time series analysis, discourse analysis, Hong Kong protests

Introduction

Critical discourse research aims to uncover the reciprocal relationships between language and power; i.e. how language forms and structures both reflect and sustain power relations (Wodak & Meyer, 2009). It has largely relied on close qualitative interpretation of texts with reference their underpinning social, political, and historical backdrops. In media contexts, however, there is growing interest in automated and quantitative approaches that can support synchronic as well as diachronic analyses of language, due to the inherent relevance of time as a variable in news (Gabrielatos & Baker, 2008; Prentice, 2010). This paper introduces, following recent related work (Smirnova, Laranetto, & Kolenda, 2017; Author, 2019), the combination of automated lexical analysis and time series modeling as a novel approach that addresses gaps in existing approaches to deepen our critical understanding of time-based discourse. The lexical analysis tool is Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010), which quantifies language use in socio-psychological categories that have particular relevance in contexts like the media. The modeling of time series data; i.e. LIWC category scores across time, applies the well-known Box-Jenkins time series methodology (Box et al., 2015). It is mostly used to analyze changes in a variable across time in order to control or forecast future values in fields like finance and engineering. Its applicability to discourse studies has only recently been explored (Koplenig, 2017; Author, 2017, 2019), the basic premise being that the temporal behavior of linguistic variables may resemble finance and engineering variables.

The approach will be illustrated by a case study of editorials on the 2019 Hong Kong protest movement in three ideologically contrastive English language news sources – *China Daily* (CD), *South China Morning Post* (SCMP), and *Hong Kong Free Press* (HKFP). In this way, a secondary objective is to show how the protests and related events were represented within a critical time span across the political spectrum. Political tensions in Hong Kong date back to its 1997 handover from the UK to China but have recently escalated over an extradition bill to legalize fugitive transfer to other jurisdictions including the Chinese mainland. Increasingly confrontational protests have since implicated other issues like democracy, alleged police/protester violence, and the future of Hong Kong as a Chinese city. The time span of interest is from 9 June 2019, the date of the largest post-1997 protest linked to the bill, to 2 October 2019, one day after the Chinese National Day by which many speculated that drastic counteractions would be taken.

In the following sections, I first discuss the background and relevant studies in the Hong Kong context including the three news sources to be compared. I then explain how the present approach can address issues with existing synchronic and diachronic analyses, leading to a deeper understanding of media representations of the protest. The methodology, results, and key theoretical implications for critical discourse research will then be presented in turn.

The Hong Kong context

The sovereignty of Hong Kong was transferred from the UK to China in 1997, ending 156 years of colonial rule and commencing its status as a special administrative region with distinct governing systems from the Chinese Mainland. Three of its most popular English language news sources occupy contrasting positions on the political spectrum. The pro-establishment *China Daily* (Hong Kong edition) was launched in 1997 by the Chinese government to present governmental perspectives for English readers. The non-profit online *Hong Kong Free Press* (HKFP), founded by independent journalists to counteract a perceived decline in press freedom, lies at the other end. *South China Morning Post* (SCMP) is the oldest English news source in Hong Kong. Its editorial stance is less clear and lies somewhere in between. After its 2016 acquisition by Chinese conglomerate Alibaba Group, it has been seen as veering towards a pro-Beijing stance while still allowing discussion of independence, self-determination, and localism (Wiebrecht, 2018). These ideological differences are expected to be reflected in differing linguistic constructions of protest-related editorials.

Post-1997 Sino-Hong Kong relations have attracted much scholarly attention across the social sciences (Cheng, 2016; Mathews et al., 2007; Ortmann, 2020). Focusing on media and language, Lee and Lin (2006) compared discursive strategies used in editorials to construct the avowed stances of the pro-establishment *Ming Pao* and pro-democracy *Apple Daily*. They found the former employed a "rhetoric of objectivity and rationality", while the latter positioned itself as the "defender of public opinion and local interests" (Lee & Lin, 2006:353). The 'umbrella movement' of 2014, most remembered for a 79-day occupation of the city centre by protestors demanding electoral reform, inspired many studies with a range of critical approaches. These include Bhatia's (2015) analysis of rhetorical tools used by SCMP (e.g. insinuation, temporal referencing, metaphor, recontextualization, reframing) to construct the movement, Mey and Ladegaard's (2015) pragmatic analysis of the 'discourse of democracy' in debates related to the movement, Lee's (2016) interrogation of selective reporting of opinion polls by different Chinese-language news sources, and Flowerdew's (2017) critical discourse historiographical analysis of the movement. The 2019 protests are likely to inspire similar research. The above studies share the common element of preferring nuanced qualitative interpretation of a limited amount of data. While this approach yields valuable insights, decisions related to data sampling and the choice of what linguistic/discursive features to focus on are not always systematically explained (Breeze, 2011). Given the avowed interest in how language reflects different political persuasions, it is also surprising that few studies have explicitly compared news sources along clearly defined linguistic variables. Moreover, since events like the umbrella movement and the current protests can quickly evolve over short time intervals, there is much room to explore how time series analyses over short spans can shed light on language changes and the attendant conceptualizations in turn.

Applying LIWC and time series modeling to critical discourse analysis

The above Hong Kong-based studies are part of a rich tradition of critical media language research. Much of this work shows how political events and relationships are construed via discursive choices that maintain divisions along gendered, ethnic, national lines, and so on. Critics often point out limited sample sizes, biased selection of features, and unsystematic analyses as its main methodological flaws. The synergy between corpus and critical methods is a response to these criticisms, as seen from the many corpus-assisted critical discourse studies in contexts ranging from politics to business (Baker et al., 2008; Koller, 2006; Partington, 2010). One of the most useful corpus-assisted analytic strategies in this regard is automated tagging with systems like the UCREL Semantic Analysis System (Archer et al., 2002). They provide bottom-up descriptions of grammatical and semantic categories that can be statistically analyzed and serve as entry points to qualitative scrutiny (e.g. Stefanowitsch & Gries, 2006). In this sense, they address all three criticisms above.

LIWC is one such system with added advantages for studies that focus on socio-psychological constructs like affect, power, and ideology (Smirnova et al., 2017). It has undergone rigorous psychometric evaluation using speaker intuitions and actual usage patterns (Pennebaker et al., 2015). In other words, its ‘bag of words’ validly and reliably reflects the underlying grammatical and semantic categories. LIWC quantifies texts under two main types of variables: ‘summary variables’ and ‘linguistic dimensions’. The four summary variables, each scored from 0 to 100, are *analytical thinking, clout, authenticity*, and *emotional tone*. They respectively compute the extent of logical vs. narrative thinking (Pennebaker et al., 2014), expertise vs. tentativeness (Kacewicz et al., 2013), personal vs. distanced (Newman et al., 2003), and positive vs. negative emotions (Cohn et al., 2004) in a text. Linguistic dimensions, the focus of this study, are normalized frequency measures of all words in a text under approximately 90 different categories. These include grammatical (e.g. pronouns, articles, prepositions, conjunctions, parts of speech, quantifiers) and socio-psychological semantic domains such as affective, cognitive, perceptual, psychological, and so on, which can reveal how the protests and related events are represented.

Furthering the above synchronic analysis, the second part of this study examines diachronic change, a growing topic of interest in critical discourse research. Corpus-assisted studies often approach diachronic change by segmenting datasets into sub-corpora each representing a particular time interval (Bamford et al., 2013; Gabrielatos & Baker, 2008), and then comparing them for statistically significant differences. More qualitative takes like the discourse-historical approach conceptualize change in more abstract ways; for example, by reconstructing the “historical interrelationships” of “thematically or/and functionally connected discourse fragments or utterances” (Reisigl, 2017:53). There are several limitations to these approaches (cf. Author, 2019). Firstly, links between linguistic/discursive units and temporal units are not always clearly articulated. Secondly, not much is said about changes over shorter time intervals that may inhere in contexts that reflect dynamic daily realities like the present editorials. Thirdly, studies tend to overlook the critical feature of interdependence or autocorrelation in time series discourse data. Common statistical techniques like log-likelihood tests assume independence between the frequencies of two aggregated time periods, but in shorter time spans we may see patterns where (near)-consecutive sessions influence one another in ways that can be modeled and critically interpreted. To address these limitations, this study will apply the Box-Jenkins approach (Box et al., 2015) to time series analysis, which uses a family of mathematical models known as ARIMA (Autoregressive Integrated Moving Average) models. Essentially, an ARIMA model expresses an observed quantity at the present time (an LIWC variable score on a particular day) in terms of quantities and/or differences between observed and predicted quantities at previous times. Linguistic variables as ARIMA models can therefore be interpreted in terms of both their synchronic content and diachronic structure, yielding deeper insights into their use across the time span. As we will see, a key theoretical implication is that determining whether linguistic variables are indeed patterned and hence 'modelable' offers a novel way to interrogate the assumption that discourse inevitably serves to construe social reality.

Data and methods

All protest-related editorials in CD, HKFP, and SCMP published from 9 June to 2 October 2019 were collected. An editorial was considered protest-related if it explicitly discusses the protests and/or related issues like the extradition bill, democracy, and so on. Reliable and exhaustive identification was relatively straightforward given the exigence of the protests. Articles published on the same day in each source were compiled into a single text file with nuisance words like author names removed. The total dataset has 201 articles and 300,081 words (CD: 92 articles/181,571 words, HKFP: 52 articles/85,165 words, SCMP: 57 articles/33,345 words). The different sizes are not inherently problematic due to the normalized nature of LIWC variables and the independent time series modeling of each series.

LIWC then computed linguistic variables in each file. This generated i) a comprehensive profile of variables for comparative analysis within and across news sources, as well as ii) each variable as a time series from 9 June to 2 October. Time series modeling was implemented by *XLStat* and all other statistical analyses and visualizations with *jamovi* (2019).

Comparative analysis

Comparison of major semantic domains

To determine the analytic focus in accordance with a data-driven approach, the first step is to compare the distribution of major semantic domains (cognitive, social, affective, perceptual, biological, drives, relativity) within and across sources (Figure 1). These domains tap into key socio-psychological constructs (Pennebaker et al., 2015) that can inform how the protests are variously conceptualized. The y-axis in all following figures show the mean proportion of words across texts in the categories on the x-axis. Error bars are 95% confidence intervals. All comparative analyses are done with factorial ANOVAs (with categories as within-subject variables and news source as the between-subjects variable). Significance level is set at p=.05.

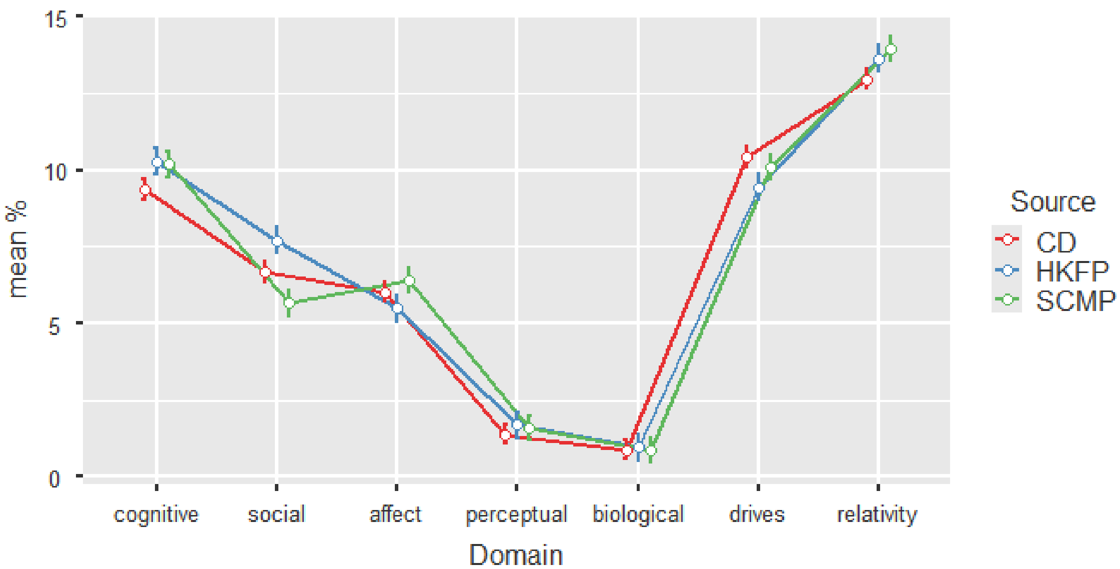

Figure 1. Semantic domains within and across sources

The domains are ranked as follows: i) relativity, ii) drives, iii) cognitive, iv) social, v) affect, vi) perceptual, vii) biological. Tukey post-hoc tests show all inter-domain differences are significant (p<.001) except drives vs. cognitive (p=1.00). Table 1, adopted from Pennebaker et al. (Pennebaker et al., 2015), show the further sub-divisions in each domain and some example words in the LIWC dictionary. Note that a word can be classified under multiple domains.

Domain	**Sub-domains and example words**
Relativity	Motion (arrive, car, go) Space (down, in) Time (end, until, season)
Drives	Affiliation (ally, friend, social) Achievement (win, success, better) Power (superior, bully) Reward (take, prize, benefit) Risk (danger, doubt)

Cognitive	Insight (think, know) Causation (because, effect) Discrepancy (should, would) Tentative (maybe, perhaps) Certainty (always, never) Differentiation (hasn’t, but, else)
Social	Family (daughter, dad, aunt) Friends (buddy, neighbour) Female/male references (girl, her, mom/boy, his, dad)
Affect	Positive emotion (love, nice, sweet) Negative emotion (hurt, ugly, nasty, hate, kill, sad)
Perceptual	See (view, saw, seen) Hear (listen, hearing) Feel (feels, touch)
Biological	Body (cheek, hands, spit) Health (clinic, flu, pill) Sexual (horny, love, incest) Ingestion (dish, eat, pizza)

Table 1. Semantic domains and example words

The top domains of relativity, drives, cognitive, and affect will be analyzed in turn below. Each constitutes at least five percent of all texts on average. The social domain is excluded because its sub-categories are relatively less comparable; ‘female/male references’ are not likely to contrast with ‘family’ and ‘friends’ in a discursively interesting way. Figure 1 also suggests that the domains have roughly similar distributions across sources, which reflects a generally similar semantic foci across the three news sources. This is expected as all three belong to the same genre of newspapers. However, interesting differences between sources become apparent when we zoom in on the sub-categories, as detailed below.

<u>Relativity words</u>

Relativity words depict movement, space, and time, and are perhaps unsurprisingly the most common category. Figure 2 shows the distribution of its sub-categories within and across sources.

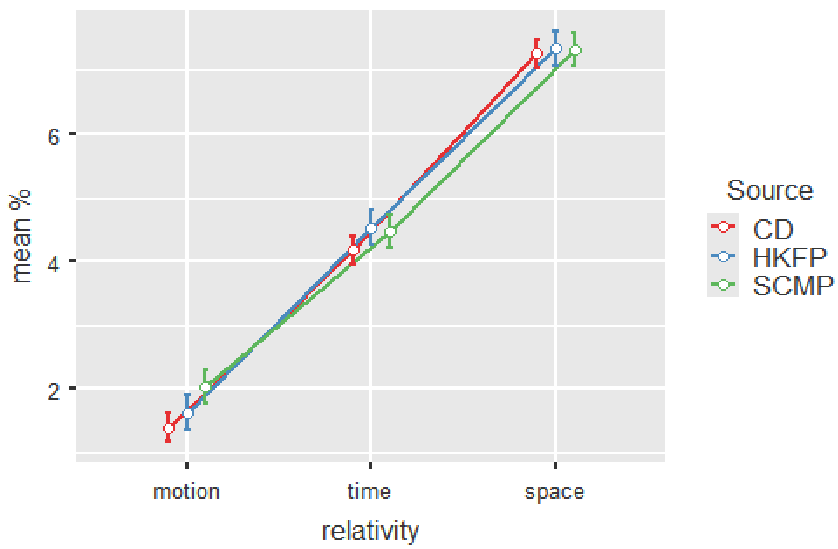

Figure 2. Relativity words within and across sources

Space words are most frequent, followed by time and motion words (p<.001) with no interaction effect (p=.071); i.e. the same order generally applies to all three sources. Relativity words might not seem interesting from a critical point of view when used in their basic spatial-temporal sense. Taking a random example from each source,

> 1. First, the government lost the power of discourse soon after the legislative process began, as the opposition hoodwinked hundreds of thousands of people into opposing the bill and joining the mass protest march on June 9 (4 July, CD)
>
> 2. On Sunday August 11, over 300 people turned out, wearing red or black, to spell out 'FREE HONG KONG' with their bodies in Taipei's Central Art Park (24 August, HKFP)
>
> 3. Coach's statement on Monday spoke of the firm respecting and supporting China's sovereignty and territorial integrity (14 August, SCMP)

The underlined relativity words in the three examples describe basic details about protest-related events, times, and places. However, cognitive linguists and critical discourse analysts have long noted that such concepts tend to be metaphorically used to conceptualize abstract concepts in subtle and systematic ways (Hart, 2011; Lakoff & Johnson, 1999). Motion and space in particular are well-known 'embodied source domains' that jointly facilitate reasoning about event structure, like in the conceptualization of political processes as 'journeys' (Chilton, 2004). In this regard, while space words are equally used by all three sources, SCMP has significantly more motion words than CD (*p*=.006). Examples 4-7 illustrate some qualitative differences in metaphorical uses of motion words from the two sources.

> 4. For a society to thrive and prosper, people need to respect and obey the law. For Hong Kong to <u>move</u> beyond the impasse and <u>turn</u> over a new leaf, the rule of law must be upheld. (29 Sept, CD)
>
> 5. Economically, the worst may be yet to <u>come</u>. Hong Kong is at risk of <u>slipping</u> into recession, as the city's Financial Secretary Paul Chan Mo-po warned on Monday. (6 Aug, CD)
>
> 6. The way <u>forward</u> is anything but clear, and deep reflection is needed to <u>bring</u> this dark chapter to an end (17 Sept, SCMP)
>
> 7. They face an uncertain future in a mature city now defined by a lack of upward <u>mobility</u>, a income gap and soaring housing prices which, together, <u>put</u> the dream of ever owning a home beyond the <u>growing</u> reach of many (4 July, SCMP)

In Examples 4 and 5 (CD), Hong Kong is conceptualized as a metaphorical mover that needs to overcome obstacles (*impasse, recession*) implied to be self-inflicted (*turn over a new leaf*). There is also a sense of passivity as the direction of movement is downwards (*slipping*), and worse situations are *coming* towards Hong Kong. The SCMP examples on the other hand describe the need for a more proactive *forward* movement to *bring* things to an end, and focus on the lack of (upward) mobility of its people in a situation that is not implied to be self-inflicted.

A limitation of LIWC shared by most other automated analyses is that it does not distinguish between literal and metaphorical uses. This may complicate the analysis of such embodied words because of their stronger tendency to be metaphorical. Determining i) the proportion of literal versus metaphorical uses, ii) the systematicity of the latter, and iii) their ostensible ideological functions would require a fuller manual analysis beyond the present scope.

Different than motion and space, time is less likely to be a metaphorical source domain because it is less experientially concrete (Lakoff & Johnson, 1999). Time words can instead be profitably analyzed with LIWC for their 'time orientation'. A past/present/future orientation is defined by past/present/future tense verbs and words that refer to past/present/future events. Previous work has examined the relationship between verb tenses and ideological messages in different genres like social media and political speeches. Djemili et al (Djemili et al., 2014) propose that ideological messages are 'timeless' and 'polychronous', giving the illusion of relevance at all times and "grouping all the temporal perspectives and cancelling them". This predicts more present than past and future tenses in ideological language. Fetzer and Bull (2012) discuss instead politicians' strategic use of past tenses to recontextualize controversial present issues and make them seem more acceptable. Figure 3 shows the distribution of time orientation within and across sources.

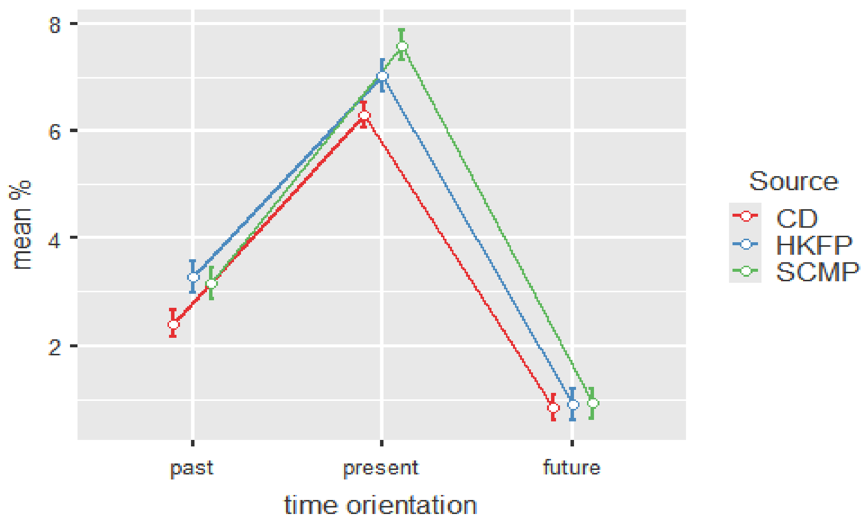

Figure 3. Time orientation within and across sources

The present orientation is most frequent followed by past and future ($p<.001$). The interaction effect is significant ($p<.01$); CD has weaker past orientation than SCMP ($p=.003$) and HKFP ($p<.001$), as well as weaker present orientation than SCMP ($p<.001$) and HKFP ($p=.005$). There were no significant differences in future orientation across all three sources, and no significant differences in past and present orientation between SCMP and HKFP. Despite this, Examples 8-10 illustrate nuanced differences in time orientation display. They show the opening paragraph of all three editorials following the formal withdrawal of the extradition bill on 4 September. This is a hallmark event in the time span that motivates editorials to reflect on the background, present situation, and way ahead (past=bold, present = underlined, future=italicized).

> 8. Here we are again with the usual suspects some well meaning, others devious and some just brain cell-challenged floundering around, suggesting ways in which the current political tension could be eased. What they don't or won't understand is the simple fact that the Chinese Communist Party is not looking for ways of lowering the temperature but is bent on defeating and humiliating the protesters. (4 Sept, HKFP)

The narrative approach in Example 8 reflects HKFP's independent press status. Its predominantly present orientation via tense and word choice (e.g. *here we are again*) emphasizes the immediacy and perpetualness of the present situation, and deems the Chinese Communist Party culpable. In line with Djemili et al's (2014) observations, past and future oriented language is consequently highly limited.

> 9. Hong Kong is *expected* to be free of any more violence, including rioting, now that the special administrative region's chief executive has withdrawn what **was** a well-intended bill to amend the extradition law. The proposed amendment **was meant** to plug the legal loopholes which have for years allowed criminal suspects from other parts of the world, including those from the Chinese Mainland, to evade justice by seeking shelter in the Hong Kong SAR. Such loopholes in Hong Kong's law have **turned** the city into a haven for fugitives, and **prevented** the SAR government from extraditing criminal suspects to other jurisdictions. (5 Sept, CD)

In contrast, Example 9 (CD) has a more balanced time orientation. Present and future oriented language is used not to convey a sense of perpetual despair but to assert a new state of freedom from violence following the bill withdrawal. Past oriented language is used to portray the bill as a well-intentioned but failed attempt to rectify legal loopholes.

> 10. At long last, the extradition bill that has embroiled the city in its worst turmoil since reunification will be formally withdrawn. Belated as it is, the decision is badly needed to take the heat out of an escalating crisis and, *hopefully*, *will* pave the way to restoring order and stability. While the change of heart by Chief Executive Carrie Lam Cheng Yuet Ngor *will* not please everyone, protesters would be wise to show similar goodwill if compromise and reconciliation are to be reached. Announcing the withdrawal in a prerecorded television address **yesterday**, the embattled leader **said** her government **was** also ready to take further steps to break the deadlock. (4 Sept, SCMP)

Example 10 (SCMP) also demonstrates a balanced time orientation. It differs from the other two in that its present and past oriented language seem less connotative, focusing more on concrete temporal details than on evaluating the state of affairs or attributing blame to either party. The future-oriented language is also less assertive than CD, hoping for rather than declaring a violence-free state.

Words referring to drives

The domain of drives is based on McClelland's (1987) Theory of Needs which claims that humans are motivated by affiliation, achievement, and power. LIWC categorizes words that refer to these and two additional subcategories of reward and risk. The relative prevalence of these words in the editorials reflects the motivational underpinnings of the protests, with specific differences shedding light on their potential links to editorial stances. Figure 4 shows the distribution within and across sources.

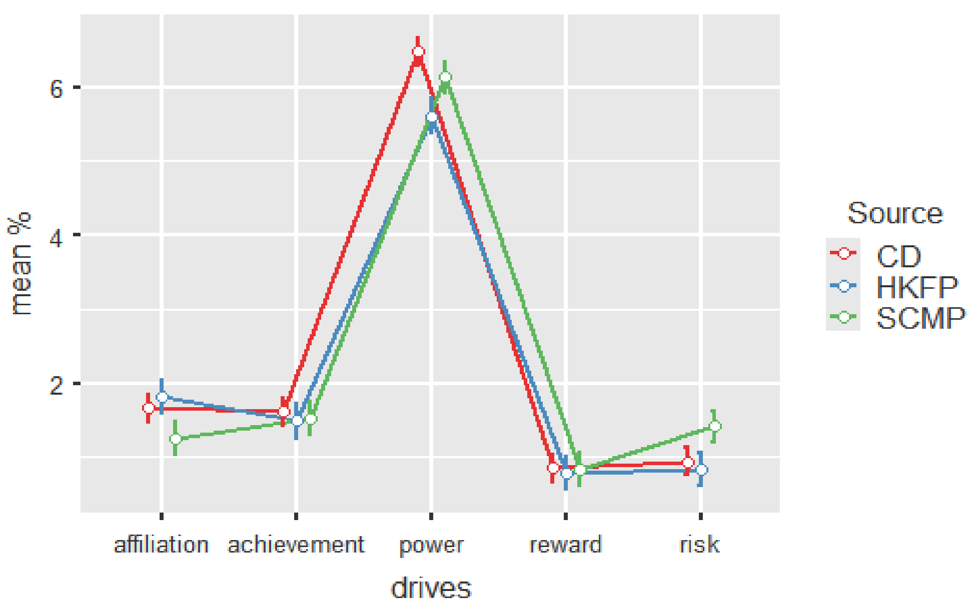

Figure 4. Drives within and across sources

Power words are by far the most frequent followed by affiliation, achievement, risk, and reward. All inter-category differences are significant (*p*<.001) except for affiliation vs. achievement (*p*=.997). While the distribution of sub-categories appears to be similar between sources, the significant interaction effect (*p*<.001) suggests interesting contrasts: i) CD has the most power words (*p*<.001) with no difference between SCMP and HKFP (*p*=.099); ii) HKFP has the most affiliation words, significantly more than SCMP (*p*=.038) but not CD (*p*=.999); iii) SCMP has the most risk words, significantly more than HKFP (*p*=.036) but not CD (*p*=.072). Each of these observations will be illustrated with reference to the three sources' editorial reflections on another hallmark event on 1 July, the Hong Kong Special Administrative Region Establishment Day. Protestors stormed, vandalized, and damaged the legislative council building.

Firstly, while power words are prevalent across all sources, this is especially the case for CD. Example 11 shows a concentration of mostly nominal labels referring to the establishment and its institutions (e.g. *government, police officers, authority, rule of law*). These institutions and 'Hong Kong society' are cast in abstract impersonal terms (van Leeuwen, 1995) and construed as expressing strong condemnation of protestor actions.

> 11.The special administrative region government and Hong Kong society strongly condemned the outrageous violence in Wan Chai and Admiralty on Monday, during which protesters charged police lines protecting the flagraising ceremony for the 22nd anniversary of the HKSAR in the morning, and then stormed the Legislative Council Complex in the afternoon and evening. More than a dozen police officers were injured. These actions were brazen challenges to government authority and the city's rule of law …If the government were to give in to their demands, which it should not, the anti-amendment forces would come up with more demands, rendering the government, the legislature and even the police force unable to operate. (1 July, CD)

Contrastively, HKFP reflecting on the same events displays a high degree of affiliation (Example 12). The emphasis is not on institutional power but on construing a sense of solidarity between HKFP and the protestors, who are more concretely represented with frequent inclusive first-person pronouns.

> 12.Victories are rare these days, and the temptation to rest on our laurels is great. After all, we have fought tooth and nail to stop a dangerous bill and this time we have tangible results to show for in contrast to the sense of empty-handedness in 2014. We deserve a pat on the back and a celebratory drink for a job well done.On the other hand, can we afford to let our guard down, however briefly? Or must we strike the iron while it's hot and keep up the pressure on our opponents until other political demands are met? (1 July, HKFP)

SCMP, on the other hand, uses the most words referencing risks; i.e. dangers, concerns, and things to avoid. Example 13 differs from the above in that it discusses an emergent mental health crisis in a way that does not appear to take either side of the political divide.

13.Sadly, three deaths with suicide notes or other references to the crisis since the mass protests began last month may not be the last linked to the controversy over the now-suspended extradition bill. Two subsequent suicide attempts went viral online, with an alarming increase in calls to counselling hotlines further raising fears of copycat behavior (4 July, SCMP)

Therefore, while power and affiliation are intuitive and observable salient opposites in editorial constructions of the protests, we also see relative preferences among news sources in terms of more ‘neutral’ categories like the evaluation of objective risks. In this regard, the sources affirm what is known about their editorial stances – CD communicates a higher degree of abstract institutional power, HKFP a concrete sense of solidarity among protestors, and SCMP a more neutral perspective.

Cognitive words

The six categories of cognitive words refer to various cognitive processes that can reveal how the editorials reason about the protests (Figure 5).

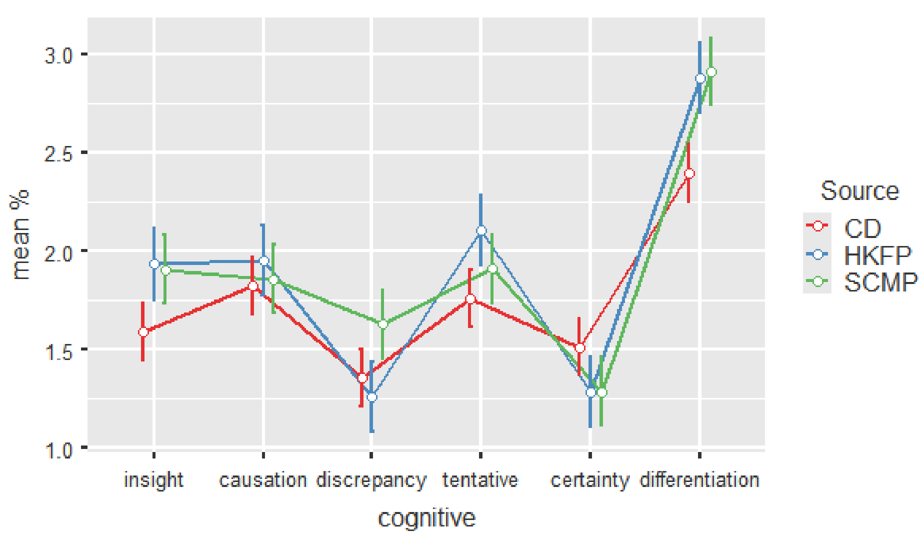

Figure 5. Cognitive words within and across sources

Similar to the previous analysis of drives, one category (differentiation) stands out as significantly more frequent than the rest (p<.001). The next three categories of tentative, insight, and causation are not significantly different from one another, but each significantly more frequent than the last two (discrepancy and certainty). The interaction effect is significant

(p<.001) but inter-source differences exist only for differentiation; CD uses the least differentiating language (p<.001) compared to HKFP and SCMP, which are alike (p=1.00).

Differentiating language includes words like *hasn't*, *but*, and *else,* which in many contexts signal disagreement or subjective distinction of entities. Examples 14-16 are editorial reflections of 21 June when protestors besieged the Hong Kong police headquarters after a peaceful sit-in. This was arguably the first major escalation since the start of the protest time span and is therefore likely to prompt rational analysis of the situation. They illustrate variation between sources in terms of their discursive construction of differences.

> 14. Although the right to peaceful protest is an integral part of a free society, people who abuse demonstrations by indulging in wanton violence that endangers the safety of others must expect to face justice …No matter what the excuse is, political or otherwise, protesters do not have a "license to break the law" in Hong Kong or any other society under the rule of law (21 and 23 June, CD)

In Example 14, CD acknowledges the right to protest but proceeds to categorically demarcate protestors from 'others', constructing the protests as a conflict between violent and non-violent individuals. This strategy of categorical differentiation is also applied to discuss the grounds (*political or otherwise*) as well as context (*Hong Kong or any other society*) of the protests, which further implies violent protestors as having uniquely political interests to cause harm to Hong Kong.

> 15. If businesspeople think that they cannot live and work in Hong Kong without the risk of a few months in gaol fighting extradition, followed by an appearance on Confessiontube and a few years in a mainland prison, then they will live and work in Singapore, and take their money and their business down there. People have a variety of views about the merits of living in Singapore, but we can all agree that it's an improvement on a mainland prison. (22 June, HKFP)

Example 15 from HKFP likewise uses differentiating language to erect discursive boundaries in a context of disagreement. The differentiation is however focused on living conditions in the two oft-compared cities of Hong Kong and Singapore, suggesting that the differences have been accentuated by Chinese influence on the former.

> 16. As reflected in the two mass protests early this month, the key to a successful protest is to stay peaceful and lawful. Confrontations that go against this principle may not necessarily win public support. These testing times make political wisdom and accountability all the more important. Regrettably, officials have shied away from facing the public, possibly out of fear that any wrong step would antagonise the situation further. But governance and public trust are at stake. (22 June, SCMP)

From this perspective, Example 16 from SCMP comes across as using differentiating language in a relatively impersonal and therefore neutral manner. The boundaries are not between concrete people and places, but abstract principles of peace versus violence, as well as the misalignment between the situation and ‘governance and public trust’. Similar to the previous analysis of words referring to drives, we see that SCMP appears to adopt a relatively neutral stance that focuses on issues rather than individuals. It should be remembered, however, that in purely quantitative terms it is CD that utilizes the least amount of differentiating words.

Affect words

Affect words describe emotions and are subdivided in LIWC into positive and negative emotion words. The linguistic representation of emotions is a fundamental topic in critical discourse studies (Mackenzie & Alba-Juez, 2019; Smirnova et al., 2017). It is also a key part of sentiment analysis in contemporary data analytics where emotions are inferred from word meanings and used to evaluate the affective properties of texts (Taboada et al., 2011). Figure 6 shows the distribution of positive and negative words within and across sources.

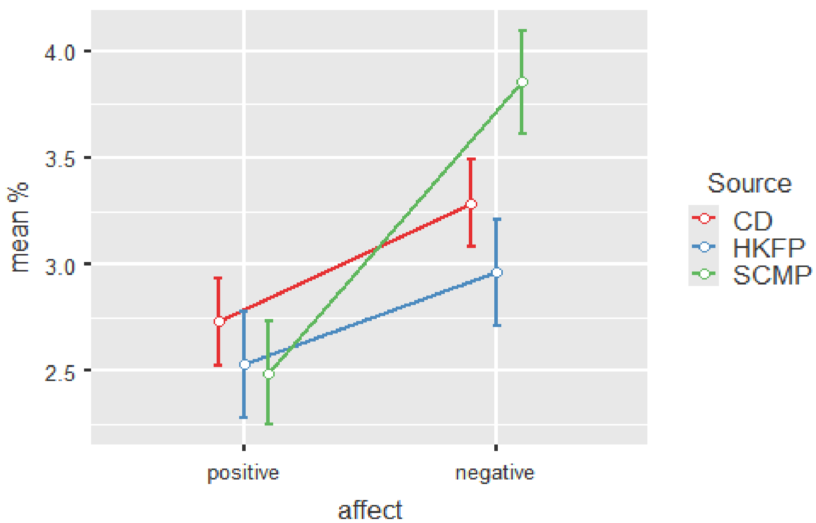

Figure 6. Affect words within and across sources

Unsurprisingly, negative words are significantly more frequent than positive words (p<.001). The interaction effect is significant (p<.001) with no difference in positive words between sources. SCMP, however, uses significantly more negative words than CD and HKFP (p<.001), which differs from the previous categories where it mostly occupied the middle position. To further probe the different sub-types of negative emotion words, we examine its specific subcategories of anger, anxiety, and sadness (Figure 7).

Figure 7. Specific negative emotions within and across sources

Anger words were significantly most frequent followed by anxiety and sadness words (p<.001), with no interaction effect (p=.14) across the sources. Protest editorials from all three sources expressed anger as the predominant negative emotion, although this tended to describe or be directed at different parties. Examples 17 to 19 are illustrative, drawing again from reflections of 1 July when protestors stormed, vandalized, and damaged the legislative council building.

> 17.Journalists covering recent protests in Hong Kong have had a difficult and, at times, dangerous duty to perform. The role they play in ensuring the public receives accurate information about the demonstrations is of great importance. But some frontline members of the media have been subjected to appalling attacks, abuse, and harassment while working. Such conduct is to be condemned (1 Jul, SCMP)

> 18.The armed wing, moreover, was maniacal, utterly committed to finishing off what they unsuccessfully attempted on June 12, when the police thwarted their earlier attack on the legislature. Perhaps the most pathetic sight of all, however, was provided by the "pan-democrat" legislators, whose role became farcical. Having earlier fanned the protests by their irrational claims about the extradition bill, they then tried to control the fanatics, only to be contemptuously pushed out of the way, with one of them, the hapless Leung Yiu-chung, even ending up sprawled on the ground (2 Jul, CD)

> 19. The perceived failure of a large scale uprising created an opening for those in power to go on the offensive. In the years since, the authorities had made damaging incursions into our freedoms and way of life, by disqualifying opposition lawmakers, banning unwanted candidates from the ballot, outlawing a localist party, expelling a defiant foreign journalist, and ceding a piece of our territory in the heart of the city to the mainland authorities. Hong Kong people were kicked while they're down, over and again. We had grown so used to these political assaults that they barely registered. Beijing had all but written us off as docile subjects who were finally beaten into submission (1 Jul, HKFP)

All three sources used anger words related to physical actions (e.g. *protests, abuse, harassment, attacks. assaults*) as well as attitudes (*maniacal, contemptuously, offensive, beaten*), to convey a generally negative sentiment towards the protests. However, SCMP (Example 17) depicts ostensibly neutral media workers as the victims of general anger, while CD (Example 18) and

HKFP (Example 19) direct this sentiment towards the protestors and establishment respectively. Similar to the previous analyses of drive and cognitive words, these observations likewise affirm the known editorial stances of the three sources. The second most frequent sub-category of anxiety words, however, has SCMP deviating again from the middle position. SCMP expresses significantly more anxiety than CD (p<.001), with no differences between SCMP and HKFP (p=.06) / CD and HKFP (p=.42). To illustrate this, Examples 20-23 are editorial reflections on the events of 23 August, the 30th anniversary of the Baltic Way as protestors held hands to form ‘human chains’ across the city. This somewhat innovative gesture prompted discussion on its symbolism and implications as editorials take stock of the protest movement.

> 20.In hindsight the city bounced back strongly from SARS, but it will not necessarily do so again when the political turbulence over the now-shelved extradition bill finally passes. A health threat that has been contained and eradicated is not to be compared with political risk and uncertainty in the minds of investors, businesspeople and talent that Hong Kong must attract if it is to be competitive in a globalised economy. (23 Aug, SCMP)

> 21.Those attending rallies throughout Taiwan aren't just standing in solidarity with those fighting for freedom across the strait, they're also collecting helmets, gas masks and other protective items for those on the streets. Likewise, the Taiwanese government can offer protection, and opportunity, to Hongkongers whose safety and freedom are threatened due to political threats. (23 Aug, HKFP)

> 22.Over the years, “human chain” rallies have become a signature tactic of independence movements around the world. Hong Kong people must guard against being misled into the slippery slope. Separatism has no future in Hong Kong. It would only bring disaster and self-destruction (23 Aug, CD)

Example 20 (SCMP) and 21 (HKFP) both contains common anxiety words like *threat, threaten*, and *uncertainty*. SCMP uses these words with reference to the non-political SARS crisis in 2003, implying that the present situation is even more uncertain and harder to eradicate. However, it focuses on business issues and is not as explicit as HKFP in attributing this threat to Hong Kong’s perceived common political enemy with Taiwan; i.e. the mainland Chinese government. Notwithstanding this difference, both sources contrast with Example 22 (CD) where no anxiety words are used. There is instead a more authoritative tone that criticizes the ‘human chains’ and calls for vigilance, short of considering it a threat worthy of anxiety. Therefore, although SCMP uses significantly more anxiety words, the ways in which they reflect known editorial stances appears to resemble the previously discussed categories.

Time series analysis

The above comparative analyses were based on the average values of each LIWC variable across the time span, giving a broad overview of differences within and between sources. From this perspective, ideological differences are reflected and perhaps indeed constructed by an aggregated observation of linguistic/discursive differences. However, an underexplored question in critical discourse research is how language use over natural time intervals, especially short-term intervals like daily news, might be patterned in ways that reveal insights into the relationship between time and the linguistic/discursive construction of reality. This perspective instead regards each linguistic/discursive variable as a time series, documenting day-to-day changes in its LIWC score. The Box-Jenkins method of time series analysis applies the following steps for each variable (Author, 2019): i) inspect and transform the series if necessary to meet statistical requirements, ii) calculate autocorrelations within the series to determine if consecutive observations are linked, iii) identify and fit candidate ARIMA models based on the autocorrelations, iv) perform diagnostic tests for goodness-of-fit, v) accept the present model if fit is adequate or find a better one. The technicalities of each step will not be elaborated here. This six-step process lead to one of two general conclusions: i) the series is randomly distributed across time such that past values have no bearing on future values, or ii) the series is patterned across time such that future values can be expressed to varying degrees of accuracy by some ARIMA model; i.e. as a function of past values.

All variables described in the above comparative analysis underwent time series modeling. Each day of publication is a time step and editorials published on the same day by each source were collectively analyzed. Most of the variables were found to be randomly distributed across the time span, also described as ‘white noise’ in statistical terminology. Figure 8 illustrates a random example of a white noise time series (cognitive words in CD).

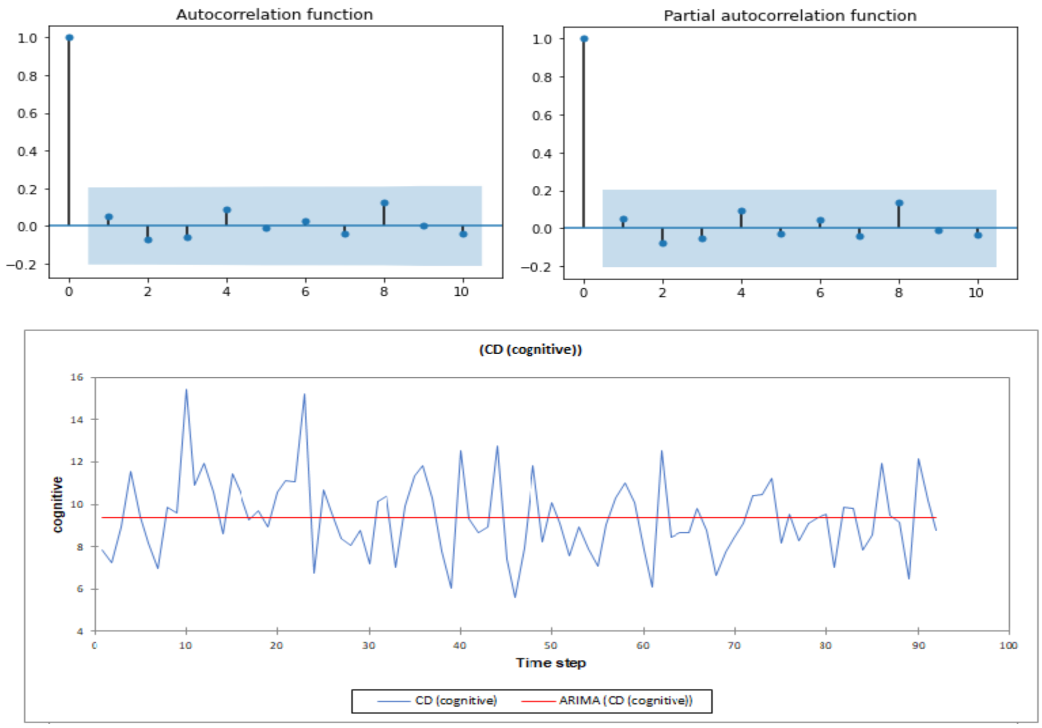

Figure 8. Example of white noise time series (CD cognitive)

The top of Figure 8 shows the (partial) autocorrelation functions for CD cognitive. The horizontal axis shows the number of ‘lags’, or daily steps apart, and the vertical axis shows the correlation coefficient for each lag. The coefficients are statistically insignificant across all lags, as visually indicated by their confinement within the 95% confidence interval region. This implies that there is zero autocorrelation in the series and hence no ARIMA models are suitable. The horizontal red line across the plotted series at the bottom of Figure 8 shows that in such cases, the best estimate at each time step is simply the mean value of the series.

Two affective variables, however, were found to fit ARIMA models. These are positive emotion words (posemo) from SCMP as well as HKFP. Figure 9 shows that for SCMP posemo the lag 1 correlation is about +0.4, which means that posemo usage over consecutive days is positively correlated - a day-to-day increase/decrease tends to be followed by another increase/decrease. This is an instance of an AR(1) model (first-order autoregressive), formally expressed as $\mathbf{y_t} = 2.4713 + 0.3925(\mathbf{y_{t-1}}) + \mathbf{a_t}$ where

- $\mathbf{y_t}$ is the value of posemo at day *t*,
- $\mathbf{y_{t-1}}$ is the value of posemo at the previous day (*t*-1)
- $\mathbf{a_t}$ is the error term inherent in any statistical model (i.e. the actual value minus the predicted value at time *t*)

In more technical detail, the model informs us that there is a stable average of 2.4713% of posemo words each day, and that a unit increase/decrease in posemo words in the previous day ($\mathbf{y_{t-1}}$) leads to a 0.3925 unit increase/decrease in the present day ($\mathbf{y_t}$). The bottom of Figure 9 shows how the values generated by this model (red line) compares with the actual values (blue line), visually suggesting an adequate fit[1]. The model can also be used to forecast tomorrow's posemo use by substituting today's values. This is a basic objective in many analytic contexts like finance, but its discourse analytic relevance is less obvious in most cases.

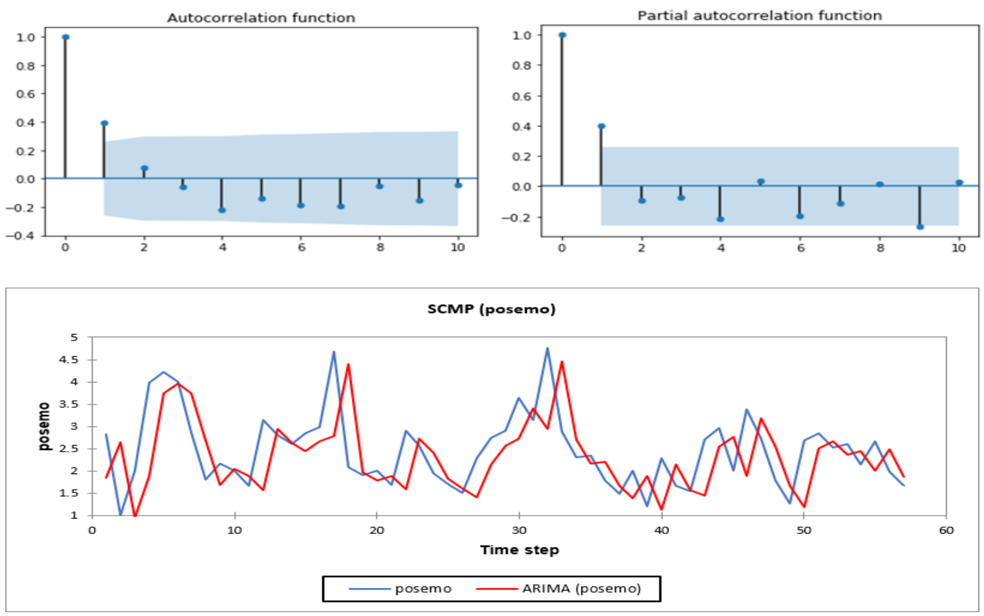

Figure 9. Observed and predicted time series for SCMP positive emotion words

The contextual interpretation of this AR(1) model is a short-term consistency or momentum in SCMP's use of positive emotion words. Values tend to increase or decrease consecutively but the time span across which this consistency most strongly holds is one day. We see that in the first half of the series, there are long stretches of consecutive rises and falls, but in the second half there are quick changes of direction after one-day spans of consistency. It appears that SCMP's portrayal of the situation is more likely to trend at the beginning with either increasing or decreasing positivity, but becomes more erratic with quickly reversing sentiments as National Day (Oct 1) approaches.

[1] There are more precise measures of accuracy such as the MAPE (Mean Average Percentage Error) of predicted vs. actual values, but they will not be elaborated here.

HKFP presents an interesting contrast as shown in Figure 10. It fits an AR(2) model (second-order autoregressive) formally expressed as $\mathbf{y_t} = 2.52 - 0.34(\mathbf{y_{t-2}}) + \mathbf{a_t}$ where

- $\mathbf{y_t}$ is the value of posemo at day *t*,
- $\mathbf{y_{t-2}}$ is the value of posemo two days prior (*t*-2)
- $\mathbf{a_t}$ is the error term inherent in any statistical model (i.e. the actual value minus the predicted value at time *t*)

This implies that associations are strongest two days apart instead of one. Additionally, the correlation here is negative; increases tend to be followed by decreases two days later, and vice-versa. In more technical detail, there is a stable average of 2.52% of posemo words each day, and that a unit increase/decrease in posemo words two days prior ($\mathbf{y_{t-2}}$) leads to a 0.34 unit decrease/increase in the present day ($\mathbf{y_t}$). This results in a sawtooth-like series across the time span as shown at the bottom of Figure 9, as directions tend to change across two-day spans.

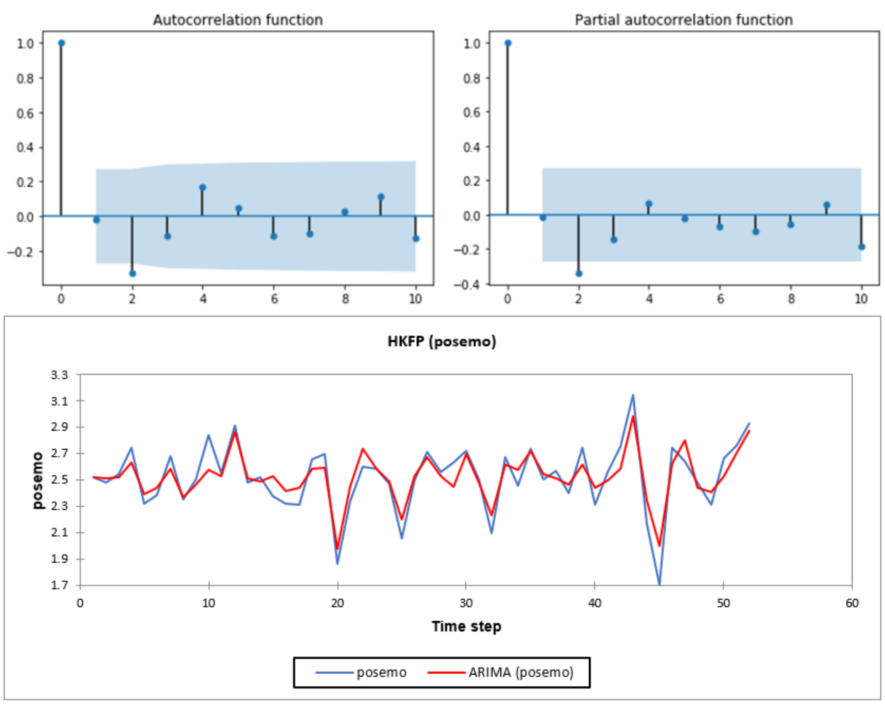

Figure 10. Observed and predicted time series for HKFP positive emotion words

Different than SCMP, the AR(2) model suggests a fluctuation in HKFP's use of positive emotion words. The two-day span is longer than SCMP but could still be considered short-term. Due to the negative correlation there is no strong linear trend in any part of the series, but only continuously volatile changes in direction across two-day spans. HKFP's portrayal is therefore more sentimentally volatile with short bursts of relative optimism followed by pessimism.

Summary

Table 2 summarizes the above comparative and time series analyses with an overall profile of the three sources.

Comparative analysis	
Relativity	• Spatial words are both literal and metaphorical. The latter conceptualizes Hong Kong along a metaphorical journey • Present orientation is most frequent • HKFP emphasizes immediacy and perpetualness of current situation in narrative style • CD more assertive of an improved situation • SCMP more hopeful and focuses more on objective temporal details
Drive	• CD has the most power words that condemn protestor actions • HKFP has the most affiliation words that depict protestor solidarity • SCMP has the most risk words that depict non-political risks
Cognitive	• CD uses the least differentiation words but tends to differentiate violent protestors from others • HKFP tends to differentiate Hong Kong and its values from other places • SCMP tends to differentiate abstract principles
Affect	• Negative emotion words are more frequent than positive ones • SCMP uses the most negative emotion words, especially anxiety • Anger is the most frequent negative emotion but expressed at different things: protestors (CD), the government (HKFP), and more general entities (SCMP) • SCMP is anxious about the protests in general, HKFP is anxious about political threats from the government, CD does not often express anxiety
Time series analysis	

Positive emotion words	• Most variables are randomly distributed across time • SCMP displays short-term consistency but becomes erratic towards National Day • HKFP displays fluctuation with short bursts of high and low emotional tone

Table 2. Summary of findings

Analyses of the four socio-psychological domains revealed broad commonalities and differences in linguistic representations of the protest movement by the three sources. All three tended to depict Hong Kong as a metaphorical mover along a journey, emphasize present orientation, and use power and negative emotion words to depict the protests as an ongoing crisis that leads to undesirable outcomes. However, a closer look at extracts shows that these representations also generally align with the known editorial stances of the sources. HKFP tends to adopt a narrative style to present the situation as urgent and anxiety-inducing, directing anger at the government and blaming it for undermining the values of Hong Kong society which protestors in solidarity are attempting to defend. CD, on the other hand, has a generally assertive tone and attributes the situation to the violent actions of protestors who are harming public interests. SCMP occupies a middle position for the most part, tending to focus on aspects of the protest that are less obviously tied to a political stance such as health, commercial interests, and abstract principles like peace versus violence. It is noteworthy that this apparent neutrality comes with the highest use of negative emotion words.

The above cross-sectional findings are enriched by considering the structural changes of linguistic variables in response to unfolding events in time. ARIMA modelling showed that most variables are randomly distributed across the time span despite their aggregated differences between sources. Clear exceptions are found for positive emotion words. While this is also random in CD, SCMP and HKFP exhibit the respective temporal patterns of short-term consistency and fluctuation. The former suggests a more stable appraisal of events in the first half of the timespan with more erratic behavior thereafter, while the latter is more volatile throughout. These patterns highlight different subjective construals of the same underlying ‘reality’ that would have been difficult to uncover by more conventional methods of discourse analysis.

Conclusion

This paper combined automated lexical and time series analysis to examine the links between linguistic choices and editorial stances over a critical time span in the 2019 Hong Kong protest movement. Different than studies that use relatively few samples or established corpus analytic approaches, it demonstrated how LIWC and time series modeling provide a coherent synchronic as well as diachronic account of linguistic representation in critical contexts. One theoretical implication is the potential of LIWC to complement other semantic taggers by focusing on socio-psychological dimensions of language. In critical contexts such as political discourse, (sub)categories like relativity, drive, cognitive and affect combine well with more content-oriented analyses at lexical and other discursive levels to explore the relationships between language and power. LIWC analysis is nevertheless limited in several ways. Firstly, as discussed earlier, it shares with other corpus analytic approaches the inability to detect figurative language like metaphor and irony. Secondly, while comparison across sources is generally unproblematic, comparisons between sub-categories of a semantic domain may be so in cases where the sizes of their respective dictionaries are too different. Thirdly, LIWC is limited to lexical analysis, and therefore cannot easily reveal higher-level discursive strategies that underlie editorial stances.

The follow-up ARIMA time series analysis also bears key implications for critical discourse research. The general complementarity between synchronic and diachronic perspectives is already well known, as outlined earlier. However, by explicitly considering potential autocorrelations in discourse data, the present method redirects attention from an aggregated view of its content to its period-by-period structure. This is most important in contexts like news, classroom, and therapy talk where dynamics over short time spans are often overlooked. As we saw in the brief technical elaboration of the two examples above, ARIMA models allow precise description of how language in neighbouring time periods inter-relate, which offers rich interpretative possibilities against the backdrop of various (critical) discourse contexts. An intriguing question would be cases where different linguistic/phenomena in different contexts (e.g. news vs. classroom vs. therapy) share similar models, thus implying previously unconsidered deep structural similarities between them.

Furthermore, the trait of 'modelability' – which distinguishes random from patterned/modelable time series – offers a fresh way to interrogate the programmatic claim, upheld by most critical discourse studies, that discourse inevitably constructs social reality. The general logic is summarized as follows. In contexts like the media, we can always assume some uncertainty or randomness in the unfolding 'social reality' that discourse aims to represent/construct. Time series modeling of relevant data (e.g. newspapers) then leads to one of two general outcomes: the variable(s) of interest are either random, or patterned in 'modelable' ways. Randomness would suggest the absence of an attempt to discursively fashion the likewise random background events, which cautions us against being too quick to conclude that discourse always presents manipulated versions of reality. On the other hand, a modelable series might indeed suggest potentially strategic or manipulative construction of (aspects) of this reality, especially when different models underpin the same events like in the present case. There is also the interesting possibility that patterned discourse does not actually construct, but merely reflect correspondingly patterned background events. A hypothetical topical example is to find some linguistic variable(s) sharing similar time series models with the daily incidence of COVID-19 they report. Either way, one should demonstrate in systematic and replicable ways if some discourse series is modelable or not, for greater confidence and clarity in claims about the relationship between discourse and social reality.

Finally, it is worth noting that time series modeling may sometimes seem to contradict 'traditional' aggregated comparisons simply because they are based on different statistical models of the data. The present case is an example where LIWC analysis using averaged scores suggests different strategic constructions among sources, whereas time series modeling suggests randomness in most variables. However, this in fact creates multiple interpretative perspectives that productively interrogate each other. Aggregated analyses that neglect temporal ordering may profile theoretical models where ideological worldviews are shaped by multiple exposure to discourse in disparate contexts (cf. Hoey, 2005), while time series analysis can deconstruct this multiple exposure and reconsider it from the perspective of temporal passage, upon which time-based discourse and social reality are inherently grounded. The ways in which these perspectives interact across various different critical discourse contexts is a promising avenue for future work.

References

Author, 2017

Author, 2019

Archer, D., Wilson, A., & Rayson, P. (2002). *Introduction to the USAS category system*.

Baker, P., Gabrielatos, C., Khosravinik, M., Mcenery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273–306.

Bamford, J., Cavalieri, S., & Diani, G. (Eds.). (2013). *Variation and Change in Spoken and Written Discourse. Perspectives from corpus linguistics*. John Benjamins.

Bhatia, A. (2015). Construction of discursive illusions in the 'Umbrella Movement''.' *Discourse and Society*, *26*(4), 407–427.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

Breeze, R. (2011). Critical discourse analysis and its critics. *Pragmatics*, *21*(4), 493–525.

Cheng, E. W. (2016). Street politics in a hybrid regime: The diffusion of political activism in post-colonial Hong Kong. *China Quarterly*, *226*, 383–406.

Chilton, P. (2004). *Analysing Political Discourse: Theory and Practice*. Routledge.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic Markers of Psychological

Change Surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693.

Djemili, S., Longhi, J., Marinica, C., Kotzinos, D., & Sarfati, G.-E. (2014). What does Twitter have to say about ideology? *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication/Social Media-Pre-Conference Workshop at Konvens 2014*.

Fetzer, A., & Bull, P. (2012). Doing leadership in political speech:Semantic processes and pragmatic inferences. *Discourse and Society*, *23*(2), 127–144. https://doi.org/10.1177/0957926511431510

Flowerdew, J. (2017). Understanding the Hong Kong Umbrella Movement: A critical discourse historiographical approach. *Discourse and Society*, *28*(5), 453–472.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics*, *36*(1), 5–38.

Hart, C. (2011). Force-interactive patterns in immigration discourse: A Cognitive Linguistic approach to CDA. *Discourse and Society*, *22*(3), 269–286.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Taylor & Francis.

Kacewicz, E., Pennebaker, J. W., Jeon, M., Graesser, A. C., & Davis, M. (2013). Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143.

Koller, V. (2006). Of critical importance: Using electronic text corpora to study metaphor in business media discourse. In A. Stefanowitsch & S. T. Gries (Eds.), *Corpus-Based Approaches to Metaphor and Metonymy* (pp. 237–266). Mouton de Gruyter.

Koplenig, A. (2017). Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digital Scholarship in the Humanities*, *32*(1), 159–168.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenges to Western Thought*. Basic Books.

Lee, F. L. F. F. (2016). Opinion polling and construction of public opinion in newspaper discourses during the Umbrella Movement. *Journal of Language and Politics*, *15*(5), 589–608.

Lee, F. L. F. F., & Lin, A. M. Y. Y. (2006). Newspaper editorial discourse and the politics of self-censorship in Hong Kong. *Discourse and Society*, *17*(3), 331–358.

Mackenzie, J. L., & Alba-Juez, L. (Eds.). (2019). *Emotion in Discourse*. John Benjamins.

Mathews, G., Ma, E., & Lui, T. L. (2007). *Hong Kong, China: Learning to belong to a nation*.

Routledge.

McClelland, D. (1987). *Human Motivation*. Cambridge University Press.

Mey, J. L., & Ladegaard, H. J. (2015). Discourse, democracy and diplomacy: A pragmatic analysis of the Occupy Central movement in Hong Kong. *Word*, *61*(4), 319–334.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception From Linguistic Styles. *Personality and Social Psychology Bulletin*, *29*(1901), 665–675.

Ortmann, S. (2020). Hong Kong's Constructive Identity and Political Participation: Resisting China's Blind Nationalism. *Asian Studies Review*, *00*(00), 1–19.

Partington, A. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora*, *5*(2), 83–108.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS One*, *9*, 1–10.

Prentice, S. (2010). Using automated semantic tagging in Critical Discourse Analysis: A case study on Scottish independence from a Scottish nationalist perspective. *Discourse & Society*, *21*(4), 405–437.

Reisigl, M. (2017). The Discourse-Historical Approach. In J. Flowerdew & John E. Richardson (Eds.), *The Routledge Handbook of Critical Discourse Studies* (pp. 44–59). Routledge.

Smirnova, A., Laranetto, H., & Kolenda, N. (2017). Ideology through sentiment analysis: A changing perspective on Russia and Islam in NYT. *Discourse and Communication*, *11*(3), 296–313.

Stefanowitsch, A., & Gries, S. T. (Eds.). (2006). *Corpus-Based Approaches to Metaphor and Metonymy* (Issue 171). Mouton de Gruyter.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Tausczik, Y. R., & 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.

the jamovi project. (2019). *jamovi (Version 1.0.1) [Computer software]*. www.jamovi.org

van Leeuwen, T. (1995). The representation of social actors. In *Texts and Practices. Readings in*

Critical Discourse Analysis (pp. 32–70). Routledge.

Wiebrecht, F. (2018). Cultural co-orientation revisited: The case of the South China Morning Post . *Global Media and China*, *3*(1), 32–50.

Wodak, R., & Meyer, M. (2009). Critical Discourse Analysis: history, agenda, theory, and methodology. In R. Wodak & M. Meyer (Eds.), *Methods for Critical Discourse Analysis* (2nd ed., pp. 1–33). Sage.