Obama in Cuba: A ‘culturomic’ analysis of frames in digital news media content

by

Ben McArthur

B.A., Western University, 2015

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Arts

in the School for International Studies Faculty of Arts and Social Sciences

© Ben McArthur 2019

SIMON FRASER UNIVERSITY

Spring 2019

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Name: Ben McArthur

Degree: Master of Arts

Title: Obama in Cuba: A ‘culturomic’ analysis of frames in digital news media content

Examining Committee: Chair: Chris Gibson
Assistant Professor

Gerardo Otero
Senior Supervisor
Professor

Leslie Armijo
Co-Senior Supervisor
Limited Term Associate Professor

Alec Dawson
Supervisor
Associate Professor
University at Albany-SUNY

Katherine Reilly
External Examiner
Associate Professor
School of Communications

Date Defended/Approved: April 16, 2019
Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016
Abstract

This study uses computational methods associated with the growing field of ‘culturomics’ to examine frames produced by digital news media sources during American President Barack Obama’s visit to Cuba in March of 2016. Using a corpus of 192 newspaper articles, ngram frequencies, topic models, and semantic characteristics (sentiment polarity and subjectivity) are assessed to identify prospective frames. Articles published by Cuba’s Granma and the American New York Times are selected as cases for comparison to represent how differences in frames manifest across media environments. Results of this analysis highlight differences in how Obama’s visit to Cuba was framed in Granma, the Times, and across the remainder of the sample. These results affirm the utility of computational methods in the study of frames, as well as other aspects of digital media content.

Keywords:  U.S.-Cuba Relations; Media Framing; Culturomics; Natural Language Processing; Content Analysis
# Table of Contents

Approval Page ........................................................................................................ ii
Ethics Statement ..................................................................................................... iii
Abstract .................................................................................................................. iv
Table of Contents ................................................................................................... v
List of Figures ......................................................................................................... vi

1. Introduction ........................................................................................................ 1

2. Literature Review ............................................................................................... 11

3. Historical Contexts, Data Sources, and Method ............................................. 19

4. Analysis .............................................................................................................. 29

5. Discussion .......................................................................................................... 46

6. Conclusions ....................................................................................................... 61

References ............................................................................................................. 70
List of Figures

Figure 1 - Normalized frequencies of ‘obama’ and ‘castro’ across corpora ..................28
Figure 2 - Normalized frequencies for variations of ‘Castro’ .............................................29
Figure 3 - Normalized frequencies for variations of ‘Obama’ ..............................................29
Figure 4 - Normalized frequencies for variations on ‘human rights’ .................................30
Figure 5 - Normalized frequencies for selected ngrams unique to NYT or Granma ..........31
Figure 6 - Normalized frequencies for ngrams referencing baseball .................................31
Figure 7 – Category-topic associations for the NYT, Granma, and general sample ...........33
Figure 8 - Inter-topic distance maps for NYT, Granma, and the general sample .............38
Figure 9 - Sentiment polarity distribution of articles within the general sample ..............39
Figure 10 - Subjectivity distribution of articles within the general sample ....................40
Figure 11 - Sentiment polarity distribution of articles within the NYT sample .................41
Figure 12 - Subjectivity distribution of articles within the NYT sample ..........................41
Figure 13 - Sentiment polarity distribution of articles within the Granma sample ...........42
Figure 14 - Subjectivity distribution of articles within the Granma sample .....................43
1. Introduction

1.1 Rationale

When considering the role of digital news media in issue formation and the direction of public discourse, the identification and interpretation of ‘frames’ remains one of scholars’ most powerful theoretical tools. Much attention has been paid to the connection between media discourse and politics, particularly with respect to news sources’ agenda-setting and priming functions, as the role of media organizations in forming public attitudes and political preferences is central to any understanding of modern political dynamics. This statement is particularly true when considering the effect of regime type on media discourse, and the extent to which a given state may co-opt news outlets to serve their own goals (Stier, 2015).

To offer some insight regarding the differences in framing practices between democratic and authoritarian political contexts, this study examines digital news media sources from Cuba, the United States, and various other nations. More specifically, it considers how frames are constructed in relation to a specific event—American President Barrack Obama’s historic visit to Cuba in March of 2016. In doing so, it seeks to answer one research question related to the substantive qualities of this coverage and one related to the methodological efficacy of a specific set of analytic techniques. First, what information regarding the frames present in coverage of Obama’s visit to the island can be extracted through the computational analysis of media content? Second, can computational methods associated with large-n studies be applied effectively on a smaller scale?

In general, the practice of framing involves increasing the salience of selected topics, or particular aspects of those topics, in news coverage for one or more reasons. Though the motivations for engaging in this behaviour are highly varied and often difficult to identify, frames themselves are comparatively accessible through the analysis of media content. New approaches to content analysis have emerged in recent years, with those relying on computation among the most innovative and widely discussed. Though their synergy is apparent to those with a background in computer science, data science, or natural language processing (NLP), applications of quantitative analysis to digital media
content remain under-represented. This under-representation could stem from the comparative ease of accessing literary textual data through the Google Books database, or social media data through API requests. Broadly speaking, however, the use of computational techniques in the social sciences has become more widespread over the past two decades, as expanding quantities of data and improvements to the quality and accessibility of software tools have created new opportunities to explore previously elusive questions. Within the social sciences and humanities, the previous two decades have produced ground-breaking scholarship in data-intensive subject areas such as network analysis (Albert and Barabási, 2002; Watts, 2007; Borgatti et al., 2009), linguistics and lexicology (Fitch, 2007; Michel et al., 2011), and content analysis (Leetaru, 2011; Shook et al., 2012).

Studies that rely on computation follow the premise that quantitative methods applied to large volumes of data can yield results that would otherwise fall outside of researchers’ capacities. While the human brain is clearly a tremendously useful analytic tool, algorithms can iterate over a particular dataset many times, and can, in the process, connect individual patterns to one another—doing so free of subconscious biases and with a much lower propensity for error. Likewise, where many hours of close reading could conceivably produce a thorough examination of one or more meaningful aspects of these data, computation can increase the depth and breadth of that examination by many degrees of magnitude. For example, the evolution of language can be monitored across millions of texts (Michel et al., 2011), hundreds of thousands of paintings can be analyzed to quantify changes in human standards of beauty (De La Rosa and Suárez, 2015), and billions of Tweets can be analyzed to monitor adverse effects of pharmaceutical drugs in real time (Bian et al., 2012). Though the use of large datasets to identify and analyze phenomena is not a new practice, each of these examples demonstrates the degree to which the boundaries of inquiry have been expanded for researchers from a variety of disciplines. This idea is articulated well by danah boyd and Kate Crawford (2012), who suggest that the increase in capacity provoked by modern computational tools “reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality” (665).
Shah et al. (2015) define computational social science by “(1) the use of large, complex datasets…(2) the frequent involvement of “naturally occurring” social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability [of their results] to social theory” (7). The authors’ first and second criteria are particularly important to this study—the first emphasizes the common belief that computational techniques must be applied to very large sets of data for optimal performance, which is tested here. In most quantitative content analyses, this would be interpreted to mean that hundreds, thousands, or millions of documents are required to obtain meaningful results. I argue that specific techniques can be successfully applied to much smaller corpora, particularly when conducting exploratory research to provide a foundation for more expansive scholarship. Rather than identifying significant characteristics of a corpus and subsequently exploring that significance through quantitative testing, this study introduces a method for asking broad questions of a dataset through computation, which can be followed with more informed qualitative inquiry.

Second, the authors emphasize the “naturally occurring” aspect of computational social science source material as a departure from the interventional practices of surveys and interviews that have defined much of social science research in the past. This is a perspective commonly associated with documentary research, which has waned somewhat in its popularity but has been demonstrated to hold significant utility—particularly in examining the role of institutions in modern society (McCulloch, 2004). By taking a computational approach to the analysis of culture, researchers are seeking to interact with the raw products of human cultural production and communication. Herein lies the synergy between computation and content analysis, a subject area that seeks to make sense of raw information in the form of textual data. Content analysts, whose research has often relied on the close reading and coding of text based on defined variables, the automation of these processes has begun to return significance in textual data that was difficult to detect through conventional qualitative interpretation.

There are several reasons to involve computation in the analysis of textual data. First, the use of computational techniques in content analysis has exponentially expanded the quantity of textual data that can be processed; automating processes that would
previously have required constant human engagement. A single researcher can now work with hundreds of millions of published works within one study, a quantity of source material that could not be consumed manually within a reasonable period of time. As the common statistical principle dictates, larger sample sizes tend to produce significant results. This principle—combined with increasing access and data storage capabilities—has given rise to so-called ‘big data’ research, whose offshoots can be found in an increasing number of academic disciplines. Researchers have also benefitted from increasingly effective digital tools and techniques. Practical integrated applications for textual analysis (e.g. Voyant Tools) have been developed in recent years, and many social science researchers with little formal training in computer science or programming have begun to write code as a part of their research. These tools allow researchers to engage with methods like word frequency distribution and lexicon-building with relative ease and efficiency.

Second, the application of computational techniques to content analysis has allowed for the quantitative measurement of abstract semantic qualities in a text, including sentiment polarity and subjectivity. While researchers have always been able to make persuasive qualitative assessments of these characteristics in a text, the development of algorithms for the calculation of sentiment polarity and subjectivity have provided a means for identifying more extensive measures of semantic elements. The emerging field of machine learning has also contributed to natural language processing, as studies from Hillard et al. (2008) and Scharkow (2013) have demonstrated the ability of both supervised and unsupervised processes to classify documents based on their content. With these advantages in mind, it is understandable that many social science researchers have begun to employ computational methods in their work. Carrying relatively few barriers to entry and finding applications in both hypothesis-testing and exploratory research, they provide opportunities to reframe and expand the questions that can be asked of one’s data.

Nonetheless, the pure automation of content analysis remains problematic. As Lewis et al. (2015, pp. 312-313) point out, computers are still ill-prepared to recognize the more complex aspects of natural language (e.g. sarcasm, irony, spam content, etc.) and should therefore be combined with qualitative analytic methods in a hybrid approach.
This principle informs the design of this study, wherein source material is selected manually, analyzed using automated techniques, and then subjected to qualitative interpretation and contextualization. While the automated selection and contextualization of source material is very common in applications of computational text analysis outside academia—particularly in the business world—these applications can yield overly superficial, general, or specific insights and can lead to mistakes in the interpretation of data.

Several strong examples of the hybrid approach advocated by Lewis et al. can be found within the growing field of culturomics, defined by Michel et al. (2011) as “the application of high-throughput data collection and analysis to the study of human culture.” In other words, culturomics makes use of new digital tools to change the way we engage in cultural studies, following boyd and Crawford’s suggestion that data-driven research can change the very nature of social science research through the application of novel techniques, methods, and paradigms (2012, 665). Culturomics has grown out of improvements in access to useful data (i.e. mass digitization of print literature, the proliferation of newspaper and database APIs), as well as advances in the tools that allow researchers to parse and analyze said data (i.e. integrated software applications, wider knowledge of programming languages). Studies associated with culturomics employ novel computational techniques to identify linkages between linguistic patterns and socio-cultural phenomena through the analysis of large quantities of data. While many large culturomic studies to this point have involved the analysis of works of literature (Michel et al., 2011; Gao et al., 2012; Caruana-Galizia, 2015), this project will rely on a corpus composed of news reports—specifically, the headlines and body text of articles published online. While this method has been employed to great effect on a number of occasions (Leetaru, 2011; Ewing et al., 2013; Dzogang et al., 2016), it remains under-utilized considering the vast and ever-expanding reserve of textual data available online.

Culturomic methods can be used to explore linguistic, semantic, and cultural questions attached to a communicating text either individually or simultaneously. The observation of metrics like usage frequency, latent topics, sentiment polarity and subjectivity in a corpus can yield interesting results when considered in context. These metrics can also be analyzed based on external variables, including geographic space and
time, as well as cultural and political contexts. It is argued here that the combination of such quantitative and qualitative techniques can provide a deep understanding of the way linguistic patterns manifest in a corpus. This study uses methods employed in culturomic studies to gather, store, and analyze data produced by the news media, with the results of said analysis explained through the lens of framing theory. This paradigm posits that information gathered from media sources must be viewed as a product of the agenda and/or biases of those who produce it, given the media’s role as a gatekeeper in the distribution of information to the public. Those who write and edit stories have significant influence over the ways in which particular events and issues are represented and can selectively elevate the position of information in the minds of their audience by increasing their salience in a publication’s coverage. This is the process of framing, which can be used to advance a particular perspective on an issue or event being discussed.

1.2 Purpose

Given the wide array of variables that affect the behaviour of media organizations and their employees, opportunities for the comparative analysis of framing practices are abundant. Frames may be considered based on their connection to broad trends, or to more narrow aspects of the context in which they are produced. This study follows the latter method, using the concept of media frames to assess coverage around a particular event. I take an exploratory and inductive approach to frame analysis, locating my findings in the historical and political context that surrounds U.S.-Cuba relations. Culturomic techniques are be employed to identify frames through the measurement of ngram and topic salience, as well as the identification of semantic characteristics in coverage.

The study’s secondary objective is to assess the usefulness of these techniques in the study of media framing on a small scale, and to offer some insight regarding potential applications in future research. As a preliminary approach to the analysis of coverage around the U.S.-Cuba relationship through hybrid computational techniques, it considers the ways in which culturomic methods might be applied in other comparative analyses of media frames, particularly across political, economic, cultural, and linguistic contexts. It
is clear that this methodology remains understudied and under-applied in the field of content analysis, and it is therefore important to assess the virtues and drawbacks of combining culturomics with framing theory.

The project draws data from a wide sampling of digital news content, but will focus on precise aspects of coverage within 192 digital news articles. It takes content from the digital English-language editions of both New York Times and Granma as specific cases for comparison with the other publications in this sample. Both of these papers sit near the top of their respective countries’ circulation rankings, and both offered in-depth coverage of President Obama’s visit to Cuba in their online editions. Beyond these characteristics, the comparison of the Times and Granma is justified by several factors related to their editorial strategies and their presumptive audiences. First, demographic data suggest a reasonable likelihood that the two publications share portions of their respective audiences. Second, the Times’ historical support of American interventionism makes it a useful counterpoint for the vehemently anti-imperialist Granma, with its close relationship to Cuba’s Communist Party. To support these assertions, it is necessary to examine several variables related to the publications’ operations, circulation, and audience demographics.

Despite the relative lack of consistent data around Granma’s audience demographics, several factors suggest that the two publications may draw similar groups of readers in their coverage of Cuba. Pew Research Center’s “Political Polarization and Media Habits” report provides data discussing the political preferences of American news audiences, including those who regularly consume content from the New York Times (Mitchell et al., 2014). It finds that readers whose political views are characterized as “consistently liberal” select the Times as one of the three most widely consumed sources of news (4). These readers also display a preference for obtaining news from multiple sources, where those identified as “consistently conservative” are much more likely to rely on far fewer sources of news (5-6). Given these readers’ proclivity for left-leaning coverage and their regular consumption of multiple news sources, it stands to reason that those with a particular interest in coverage of Cuba could be among those accessing the English-language edition of Granma.
While there is a lack of similar empirical data regarding the demographic or ideological features of Granma’s readership, several of the publication’s fundamental characteristics support the notion that it may attract Anglophone readers with similar attributes to Times readers. First, it is highly unlikely that any significant proportion of Granma’s digital audience is made up of Cubans living in Cuba—stable and consistent Internet access on the island continues to be prohibitively expensive for most Cubans, and logic dictates that the citizens of a majority Spanish-speaking nation would be unlikely to use this limited amount of access to read the English-Language version of a paper that is widely available printed in their native language (Baron and Hall, 2015: 343). Furthermore, Granma is rarely seen as a source of truly objective information, both within and outside of the country. With these factors in mind, the demographics who are likely to consume Granma’s English-language digital edition can be narrowed to two principal categories: Anglophone readers with far-left political views, who may be more inclined to accept the information presented by Granma as objective, unbiased, or as a counterpoint to American corporate media coverage; second, Anglophone readers that are aware of the publication’s role as a mouthpiece for the Cuban state, but whose habits related to the consumption of multiple news sources have led them to use the paper as a proxy for understanding the views of Cuba’s government.

With these data in mind, I argue that the Times is among those American publications that are most likely to share readers with Granma in spite of the profound differences between the two with respect to their economic structures and journalistic models. In the absence of strong empirical data supporting this assertion, the comparison of Granma and the Times is predicated on the notion that individuals who identify as members of the political left and lack Spanish language skills necessary to read the principal digital edition of the Cuban paper can be identified as prospective audience members.

The second factor that informs the selection of Granma and the Times as cases for close examination is related to the two outlets’ editorial behaviours, particularly their divergent perspectives on the issue of American interventionism. Many outside of Cuba have argued that American corporate media organizations display a tendency to cover issues based on their implications for the foreign policy interests of the United States,
with Herman and Chomsky’s (1988) *Manufactured Consent: The Political Economy of the Mass Media* underpinning much of the discussion around the topic. The authors propose several operative principles that undermine the legitimacy of American mass media as a source of objective information, among them a latent anti-Communist perspective that informs corporate media discussion of issues related to conflict between the United States and others. Krishnaiah et al. (1993) cite the *New York Times*’ coverage of the Soviet intervention in Afghanistan as strong evidence in support of this framework, as they find a direct connection between how “threatening” a particular issue or events is perceived to be according to the erstwhile foreign policy and the relative favourability of the publication’s coverage. The authors find that “news coverage was favorable toward elements defined as non-threatening based on U.S. foreign policy and unfavorable toward threatening elements”, though they observe a slight increase in the favourability of coverage related to threatening elements after 1985 (652). A review of *Times* editorials from February 2019-March 2019 suggests that these characteristics have persisted, alongside the observed trend toward more mixed tonal attributes in the paper’s coverage. In this period, the *Times* Opinion Section published an array of pieces contemplating the various ways in which the United States and others can participate in the resolution of Venezuela’s ongoing political turmoil. Though few of these editorials advocate for military intervention in Venezuela, they all engage with the various ‘soft power’ options available to those nations who may wish to exert their influence over the country’s affairs (Board, “Venezuela’s Border Standoff”, 2019; Board, “Venezuela’s Hunger Games”, 2019; Castañeda, 2019; Corrales, 2019; Kronick, 2019; Rogers, 2019).

There is evidence from inside Cuba that this perception of the *New York Times* and other American media organizations persists among members of the country’s intellectual class, in spite of the recent move toward normalization. Yoan Karell Acosta González (2017) of the University of Havana presents a critical discourse analysis of *Times* coverage in the year following the new U.S. Cuba policy was announced, which highlights a number of changes in the Cuban state’s view of *Times* coverage over the past several decades. While the author highlights an increase in the frequency with which bilateralism and normalization are referenced in the publication, she suggests that advocacy for American intervention in the affairs of Cuba is particularly salient within
the *Times* coverage. This frame, the author muses, is “naturally pushed forward as if it were an irrefutable truth” (185). This statement, presented as part of a critique that makes no secret of its support for Cuba’s political and economic model, supports the notion that the *New York Times* continues to be perceived by the Cuban political establishment as explicitly interventionist in its coverage of Cuba.

*Granma*, by contrast, represents a clear counterpoint to interventionist coverage based on the publication’s proximity to the vocally “anti-imperialist” Cuban state. As an extension of Marxist-Leninist dogma, “anti-imperialism” has come to signify a broad sense of opposition to American involvement in the political, economic, and cultural affairs of Cuba, and is prevalent in *Granma*’s digital, English-language coverage over the past several years (Rabkin, 1992). Though the concept of anti-imperialism is not commonly evoked to refer to military intervention, particularly in recent times, it can be helpful in explaining the Cuban government’s firm stance on bilateral relations between the island and their American neighbours. Where González’ suggests that the United States’ right to involve itself in Cuban affairs is presented as “irrefutable truth” in the *Times*, I observe a similar assumption of irrefutability in *Granma*’s discussions of American imperialism in the context of its relations with Cuba (Morales, 2016; Lissy Rodríguez, 2017; Sergio Alejandro Gómez, 2018).

The comparison of *Granma* and the *New York Times* can therefore be justified based on probable overlap among their respective audiences, as well as their diametric opposition on the issue of American state involvement in the affairs of Cuba. While the differences between *Granma* and a more ideologically conservative American publication would likely be much more clearly defined, I argue that the *Times* is among the most likely to share readers with the Cuban paper. Though the effect of differences in framing on audiences is not examined here, the comparison of these two publications with a general sample in relation to the construction of frames serves as a starting point should more comprehensive data become available. Further, the radical differences in each publication’s identity and context provides an ample basis for evaluating culturomics’ ability to identify differences in framing through the examination of content. By considering one case wherein the outlet in question operates based on the symbolic principles of independence and truthful reporting, and another which does much of its
work at the behest of the Cuban state, the likelihood that there will be observable differences in the construction of content is high. The extent to which these differences can be described through culturomic analysis should therefore provide a useful baseline estimation of the techniques’ usefulness. In future examinations of cases with more subtle differentiation, this understanding of culturomics’ potential for application will aid in the development of suitable methodologies.

There are several reasons that President Obama’s visit to Cuba is a relevant case for study. The first is related to the historical significance of the event itself—it represents a landmark shift in the diplomatic relationship between the two countries, with their respective governments having severed ties amidst the Cuban Missile Crisis of 1961. Obama was the first American President to visit Cuba in 88 years, and would meet with Cuban President Raul Castro, brother of Fidel—the two most symbolic participants in the overthrow of Fulgencio Batista’s pro-American regime and the architects of post-revolution Cuban society. Debate over their actions has remained at the fore of public discussion in both Cuba and the United States, as multiple waves of migration from Cuba to the United States have produced a large and politically active expatriate population.

Obama’s visit to Cuba is also particularly useful in comparing how each of the two papers operate. The visit resulted in similar levels of media scrutiny in both nations—levels that far exceed any other moment related to their relationship in the 10 years preceding. While there was no shortage of media discussion around U.S.-Cuba relations prior to the American President’s visit, this occasion represented a rare example of the two countries sharing ‘front-page’ news. With these characteristics in mind, Obama’s visit to Cuba represents a novel opportunity to advance academic understanding of how the U.S.-Cuba relationship is framed in digital news content, while simultaneously providing an opportunity to assess the utility of a culturomic methodology in pursuit of this insight.

2. Literature Review

2.1 ’Framing’ in News Media

This study relies on framing theory to evaluate the results of its analysis of news articles. Framing emerges as a concept with Erving Goffman’s Frame Analysis: An essay on the organization of experience (1974), which presents framing as a cognitive
behaviour that allows individuals to organize and interpret information internally. This concept grows to describe the organization of experience as an internal or external act, and is subsequently applied to the study of mass communications. Divergent perspectives on the theory are synthesized by Entman (1993), who defines framing as a process that necessarily involves selection and salience, both of which are employed to “promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation” (52). Entman’s identification of the terms by which topics are framed is especially important, as it sets wide boundaries for what one might consider framing. To frame a topic is not only to martial the forces of selection and salience, but to do so based on one of the motivations outlined above. This definition also represents a slight departure from the concept described by Goffman, as it incorporates intent as a variable that must be factored into any understanding of the framing process.

Entman identifies “the communicator, the text, the receiver, and the culture” as the four contexts in which framing occurs (52). In doing so, he differentiates between internal frames (schemata) that produce conscious or unconscious biases in the communicator, and those frames that organize ideas during the process of communication. Communicators use their internal schemata to make decisions about how they will present a topic to their audience. The text bears these decisions in the form of frames, which are communicated through the use of language, imagery, and external reference points. The receiver processes these frames, filtering the information through their own schemata. Finally, the culture is the lexicon of frames that exist in a particular social context—both within the schemata of receivers and in the accumulation of texts that are prevalent in that context.

Given its focus on the relationship between media organizations, political contexts, and news content, this study is primarily concerned with the roles of the communicator and the text. While it does not prioritize frames’ effect on the receiver, it does offer limited discussion of these frames’ position in the culture that surrounds news coverage in Cuba and the United States. To engage with these concepts, the notion of salience as identified by Entman is particularly important. Salience represents a link between the quantitative and qualitative approaches taken in this study’s analysis, as the assessment of particular terms’ and topics’ salience through computation provides the
fundamental basis upon which frames are deemed to be present in the corpora under consideration.

In the news media, framing occurs in multiple locations. The work of reporters, editors, commentators and publishers necessarily involves the use of framing, both in the way topics are interpreted through individuals’ schemata and in the way they are used in presenting information to an audience. Newsmakers’ decisions regarding what to publish and how to cover particular topics can themselves be viewed as the product of schemata, which have significant influence over supposedly rational or logical processes of choice (Kahneman and Tversky, 1981, 1984). The process of making news can therefore be viewed as a constructive one, the product of various social and institutional interactions rather than the inevitable coverage of what is already known to be most important to the public (Tuchman, 1978; Gans, 1979; Fishman, 1980). Acknowledging the social constructive nature of media discourse—as well as other forms of written communication and documentation—raises important questions regarding the relative objectivity of news discourse. Media actors do not portray events based on their objective qualities, but rather based on imposed frames communicated through the precise and intentional use of language (White, 1987).

Accepting that the use of frames by the news media constitutes the construction of reality through communication, the content of media discourse can be seen to have significance beyond what is communicated explicitly through its language. Frames are used for different purposes by newsmakers, and are the products of different intentions and degrees of premeditation. For example, Gitlin’s (1980) study of the relationship between the Student New Left movement and the news media leads him to suggest that in many cases, frames are used by newsmakers to distill complex topics in a way that makes them digestible for a general audience. For Gitlin, framing is a practice that increases the efficiency with which information is diffused across societal networks, with newsmakers serving a key function in this distributional pathway. Framing, by this definition, is a descriptive process that is employed to clarify rather than obfuscate.

However, frames can carry equally powerful agenda-setting powers. For example, Iyengar (1991) identifies “episodic” and “thematic” frames in television news—the former referring to frames that focus on the particularities of an event and the latter
referring to frames that connect a specific event to a broader theme or context. In the political news domain, these qualities have tremendous significance. Iyengar suggests that American television news’ tendency toward using episodic frames de-links specific situations or events from the long-term forces that drive societal change, deflecting public attention away from the political actors that regulate such change. This is an example of the ways in which frames are useful to those who seek to co-opt media platforms for strategic ideological reasons, including political or economic elites, interest groups, or technocrats who may wish to selectively elevate or diminish a particular narrative in news coverage.

Accepting that framing cannot be characterized as inherently virtuous or nefarious also allows for the consideration of frames without emphasizing their intent. While it is entirely possible that, in the case of U.S.-Cuba relations, the frames used in news coverage are constructed based on a desired outcome, it is equally possible that the formation of these frames is influenced by the internal schemata of the individual(s) who produced the content in question. This aligns with the perspective advanced by Pan and Kosicki (1993), who suggest that the interests and contributions of all parties who participate in the news ecosystem can be considered in the analysis of frames and framing practices. Such a perspective allows scholars the opportunity to consider cases holistically or break them down into particular elements, leading the authors to suggest that frames in news coverage “may be studied as a strategy of constructing and processing news discourse or as a characteristic of the discourse itself” (57).

This perspective presents an opportunity to connect the analysis of frames in textual data with broader socio-economic and cultural forces, a fundamental aspect of critical discourse analysis (CDA). Van Dijk (1993) suggests that a critical analysis of discourse, irrespective of the theories or methods through which it is conducted, must engage with the exertion of power and dominance through communication. This can be observed, the author suggests, “through the enactment of dominance in text and talk in specific contexts, and more indirectly through the influence of discourse on the minds of others” (279). Norman Fairclough (2001) takes a similar view, suggesting that “seeing language as discourse and as social practice” requires one to consider “the relationship between texts, processes, and their social conditions” (26). In a discussion of framing,
this suggests that a complete understanding of frames’ role in discourse requires that these structures be analyzed in relation to the processes and contexts from which they emerge. As such, this study follows the principles of CDA outlined by Van Dijk and Fairclough, seeking to describe evidence of frames used to describe Obama’s visit to Cuba that may be connected to the broader socio-economic and cultural contexts in which the content was produced, as well as their potential to provide different points of reference for audiences.

2.2 Culturomics and the Computational Analysis of Textual Data

Beyond discussing the substantive results of frame analysis, this study considers the ways in which new methodologies can empower scholars to approach analysis of media content in novel ways. Of particular interest are the techniques associated with the emerging field of culturomics. Coined by Michel et al. (2011), culturomics refers to “the application of high-throughput data collection and analysis to the study of human culture.” This implies a paradigmatic and methodological shift in social scientists’ and humanists’ approach to studying cultural change, offering techniques and perspectives that are inextricably tied to new technologies.

In general, culturomics attempts to bridge gaps in our understanding of human culture through computation. The authors demonstrate several of the techniques associated with the field through the use of a large corpus to investigate social and cultural phenomena as reflected in written language. Their study examines a corpus composed of 5 million digitized books spanning six centuries and several languages (~4% of all books published) which has been made available through the Google Ngram Viewer. While this dataset is too large for any one human to read, modern computing power has allowed the researchers to observe trends in the lexicon, grammar, and content of written works. Michel et al. are primarily concerned with the ways a large corpus can be used to observe patterns in cultural and linguistic change. These categories include the ways in which people, settings, and events are represented in written language, as well as changes in the mechanics of language itself. To do so, the authors compute several metrics based on ngrams in their corpus—a term used a sequence of n words or figures (176).
Perhaps the most useful and accessible metric for engaging with ngrams is normalized frequency. Raw frequency refers to the total number of times that a word or sequence of words appears in a corpus, where normalized frequency is calculated by dividing the number of occurrences of an ngram in a given time period by the total number of ngrams in the corpus. This number can then be multiplied to produce an average number of occurrences per $x$ words, allowing corpora of different sizes to be compared effectively. In practical terms, the measure of normalized frequency is useful for quantifying salience of an ngram in a corpus, understanding high frequencies to be associated with more frequent mentions of a word or phrase.

While the examination of frequencies for ngrams that are shared among multiple corpora can be used to identify common characteristics, it can also illuminate important differences. For example, the comparative analysis of ngram frequencies can be used to quantify censorship in literature, which has proven empirically difficult in the past. Michel et al. offer several means of detecting censorship, plotting the frequency of names, words, and phrases across language groups. In one example, the authors observe differences in the mention of the Jewish artist Marc Chagall across English and German corpora; usage of Chagall’s name rises steadily in both languages until the years 1936-1944, during which time it almost completely disappears from the German corpus. By contrast, his name continues to increase in frequency across the English corpus, and appears a mere 100 times in the German corpus during the 8 years following the conclusion of the Second World War (1946-1954). The authors observe similar patterns in the cases of “Trotsky” (Russia), “Tiananmen Square” (China) and “the Hollywood Ten” (United States) [pp. 180-181]. Though the substantive conclusions of this analysis are significant in themselves, it also demonstrates how culturomic techniques can find meaning in the relative absence of particular characteristics in a given corpus. This principle informs the analysis conducted here, as Granma and the New York Times are compared based on ngrams that are highly salient in one corpus but absent from the other. As in the aforementioned studies, this method provides novel and meaningful insight regarding the frames present by each publication.

One important consideration for content analysts that arises from the interpretation of ngram frequencies is the possibility of mistaking linguistic change for
cultural change. Shea and Spoveri (2012) offer a strong example of how a researcher might address this issue, using the Google Ngram database to measure relative levels of ‘incivility’ in political rhetoric. The authors measure the usage frequency of several synonymous adjectives alongside “politics” (i.e. “mean politics,” “bitter politics,” “hateful politics,” “filthy politics,” and “nasty politics”) in literature over time to quantify relative levels of negativity associated with American political speech. They include this variety of adjectives to account for changes in the way ‘nastiness’ is described across time periods, acknowledging one of the most important obligations of those who engage in culturomic analysis—to ensure that the patterns they observe are the product of changes in a phenomenon itself, not the way it is described. For example, the emergence or disappearance of a relevant term in frequency analysis results may be evidence of censorship or propagandist usage, or it could simply represent a manifestation of the natural evolution of language over time. The same principle informs the design of this study’s analysis, as differences in the frequency with which particular individuals are mentioned in each corpus are qualified through the inclusion of several variations on each name under consideration.

Beyond the examination of ngram frequency, culturomic studies have employed far more advanced statistical and algorithmic methods for analysis. For example, studies concerned with the semantic characteristics of written language have proven incredibly insightful in assessing the tone of content. Tone, though admittedly broad in its significance, is generally associated with the use of particular words, phrases, imagery and reference points to convey a primarily positive or negative assessment of a topic. In content analysis, tone can be represented through the measure of sentiment polarity—wherein algorithms assign defined portions of text a score between -1 and 1, with the latter indicating positive sentiment and the former negative. Leetaru (2011) uses sentiment polarity to observe changes in the ‘tone’ of media coverage across various Middle Eastern countries. The author finds a strong correlation between levels of sentiment polarity among global media sources reporting on the countries in question and levels of political stability in these countries, leading him to suggest that culturomics may have applications in assessing and predicting changes in global political dynamics through the interpretation of media tone. Following from Leetaru’s conclusions regarding
the connection between tone and political dynamics, this project uses semantic analysis to examine differences in tone between *Granma* and the *Times*.

Though many of the studies mentioned here focus primarily on the relationship between politics and language, applications for culturomic methods straddle disciplinary lines. Gao et al. (2012) use culturomic techniques to analyze differences in the physics of long-term correlations between social and natural phenomena as represented in language. Chung et al. (2014) analyze a database of personal essays, assessing the way linguistic representations of personal concerns vary across geography. Shook et al. (2012) take a “geospatial visual analytical approach” to culturomic analysis, analyzing the complete English edition of Wikipedia to create a geographic heatmap of sentiment associated with discussion of “armed conflict” during the year 2003. While the diversity of disciplinary and theoretical foundations among culturomic studies demonstrates the relative novelty of this field, it is also a testament to the potential versatility of culturomic methods.

The perspective and techniques associated with culturomics inform this study’s analysis of frames in news media content, emphasizing the use of computational discourse analysis as an exploratory tool rather than a means of affirming qualitative results. Rather than using these methods in an effort to engage with framing as a theoretical paradigm, they serve as descriptive tools—techniques that can be used to identify unconventional variables and patterns in the application of frames. While the likelihood that such an approach will produce generalizable knowledge is low, the diffusion of computational techniques across the social sciences and humanities is itself valuable.

In order to understand the ways in which Barrack Obama’s visit to Cuba was portrayed by Cuban, American, and global news media, it is crucial to examine the linguistic devices that construct frames. However, given the relatively short timeframe from which this project’s data was sourced, it is difficult to observe any significant shifts in grammar or lexicon with respect to media coverage of Cuba. Rather, this project is interested in the media’s emphasis of particular themes, terms, or specific issues in relation to a single event. This information serves as a broad representation of media discourses around Obama’s visit, particularly within Cuba and the United States, and
offers some opportunities to connect the framing of this event with long-present themes in these discourses.

3. Historical Contexts, Data Sources, and Method

Before discussing this study’s analysis of news articles, it is necessary to describe the ways in which data were collected, stored, and processed, as well as the general conditions of the contexts in which they were produced. First, the structural qualities of both the Cuban and American news ecosystems are described in terms of their political, legal, and economic characteristics. This information is included to offer various points of reference during the interpretation of the study’s analysis, which produces several results that are directly connected to the media environments in which Granma and the New York Times operate. Both the production and consumption of news content occur in radically different ways across Cuba and the United States’ legal, political, and economic contexts. Therefore, an understanding of the differences between these environments is crucial to determine how frames are constructed in both countries. Second, the sources from which data was obtained and the methods used to analyze the data will be described in detail, alongside a brief justification of their selection and application. While many of the methods employed in the study’s analysis involve computation, they are supported by many conventional qualitative techniques to create a hybrid approach.

3.1 Characteristics of Media Production and Distribution in the United States

In the United States, media organizations enjoy various legal protections for their reporting and publishing activities. Press freedom and freedom of speech are enshrined in the First Amendment to the country’s Constitution, and are widely considered to be central tenets of the American democratic model (US Const. amend. I). In general, these rights guarantee that media organizations can report or publish whatever they choose so long as it is not libelous, and that writers and editors are not subject to undue state intervention in their activities. Such fundamental protections for press independence follow the principle that an independent press is essential to creating a well-informed public through the distribution of high-quality information. This public, in turn, will be
sufficiently informed of their country’s affairs to participate effectively in its democratic process.

Despite strong constitutional protections for reporters and publishers, market forces and the advent of new technologies have also changed the production and distribution of journalistic content in the United States quite dramatically. According to Pew Research Center (2018), American newspapers have seen consistent decreases in circulation and advertising revenues since the early 2000s with the rise of the Internet and emergence of social media, though digital ad revenue is increasing across the industry. In spite of industry-wide decreases in newspaper circulation, many American news consumers are migrating to digital news sources—the New York Times, for example, reported a 42% increase in digital circulation in 2017.

While this mass migration to digital news content is economically significant, it also has important implications for the actual content of American newspapers. Where early online versions of newspapers would often be replicates of print editions, modern digital editions are often differentiated from those in print. Several factors have informed this trend toward unique digital content, though the increasing centrality of social media to the distribution of content is particularly significant. The Times, which has long been one of the most widely circulated national newspapers in the United States, has come to be symbolic of many large-scale trends in American digital news. In 2014, an internal review of the newspaper’s digital strategy was leaked to the press, containing several revelatory insights into the paper’s changing business and editorial strategies. For example, the document revealed that fewer readers than ever before were visiting the Times homepage, and were much more commonly discovering content through sharing across social media platforms (New York Times Innovation Report, 2014). This and other insights led internal reviewers to emphasize a need for deeper collaboration between the business staff and newsroom staff, particularly in audience development and the “packaging” of information. The recommendations contained in the Times report also make reference to similar trends across other prominent media outlets, including the Washington Post, Yahoo News, and the Wall Street Journal (Innovation Report, pp. 23-25). In addition to its implications for audience development practices, this statement suggests an increasing differentiation of editorial strategy between print and digital
editions of papers like the Times, and affirms the notion that digital editions of prominent newspapers merit empirical study. Scholarship of the agenda-setting function of media has also provided evidence of this phenomenon dating back to the early years of digital newspaper editions—for example, Althaus and Tewksbury’s (2002) study finds that online and print readers of the Times adjusted their respective agendas differently to one another. While I do not explore the differences in agenda-setting outcomes between print and digital formats, further research related to framing of the U.S.-Cuba relationship would be well served by examining this topic in more depth.

3.2 Characteristics of Media Production and Distribution in Cuba

Where press freedom retains a sense of symbolic importance in the American democratic model, the Cuban state has pursued an entirely different path in its relationship to the media. Even among countries with authoritarian leadership, the extent to which the Cuban government exercises control over the content and distribution of news media is unique (Walker and Orttung, 2014, p. 72). It is one of only a few nations in which the national government seeks total control of mass media—that is, the state seeks to control the production and distribution of news rather than simply regulating independent sources’ content. As a result, the country currently ranks 172nd of 180 countries on the nonprofit group Reporters Without Borders (RWB)’s Press Freedom Index, meaning it is considered one of the world’s least free media environments (2018).

As reports from Freedom House (2018), RWB (2018), and the Committee to Protect Journalists (CPJ) [Lauria, 2016] point out, private media organizations are still formally banned under Cuba’s constitution. While Article 53 affirms citizens’ right to press freedom and freedom of speech, this is only tolerated if it is seen to contribute to “the objectives of socialist society” (Cuban Const., Article 53, 2002). According to the aforementioned CPJ report, the country’s penal code also contains provisions for the prosecution of individuals for defamation, insults, and contempt of authority. These concepts provide fairly broad mandates for the Cuban authorities to arrest and detain those who engage in any form of dissent—including journalism that is perceived to be critical of the Cuban state.
The politics of journalism also differ between the United States and Cuba, as the small number of traditional newspapers that operate in Cuba are directly connected to the national government. Three newspapers are circulated nationally, with each representing a particular organ of the government. According to Miyares (1999, 79), the daily newspaper *Granma* is “the voice of the Central Committee of the Cuban Communist Party and the country's newspaper of record.” It operates alongside the National Workers Confederation’s *Trabajadores* and *Juventud Rebelde*, which represents the Central Committee’s Youth organization. There are several regional papers as well, including Havana’s *Tribuna de la Habana*, Santiago de Cuba’s *Sierra Maestra*, and Holguín’s *Ahora*, among others. Given their proximity to state organizations, Cuban newspapers do not serve the same function as their American counterparts. There is little evidence of competition among these publications for readers, for example, nor do they appear to be concerned with profitability. Miyares’ acknowledgement that *Granma* and its compatriots generally act as mouthpieces for various actors within the Cuban state is therefore central to any understanding of the content produced by the publication, which tends to disseminate and reinforce state messaging rather than police the state’s activities as is the case in the American model.

Economic forces have had similar effects on the Cuban news ecosystem to those unfolding in the United States, though their media system entered a much more significant state of crisis much earlier. The severe economic downturn that occurred during the so-called “Special Period”—the years following the collapse of the Soviet Union, which provided the vast majority of funding and imports to the island—led to a drastic reduction in newspaper circulation and reporting jobs (Miyares, 1999). Cuba’s economy has improved slightly since that time, but it remains underdeveloped—particularly as it relates to information and communication technologies (ICTs).

Since the reforms of 2011—which loosened restrictions on private enterprise in the country (Franc, 2011)—Cuba has also witnessed the development of a vibrant ‘blogosphere’. This term describes a community of news sources published online that is made up of both independent and state-run operations, varying in size and readership. The bloggers Yaoni Sánchez and Elaine Diaz are two of the most high-profile, with the former receiving several international awards for her journalism and the latter earning a
Neiman Fellowship from Harvard University in 2015. Pioneering studies by Vicari (2014a, 2014b) have found that content published by Cuban bloggers is more ideologically diverse than in traditional national newspapers, often informed by criticism of state activities and personal narratives. However, Cuba’s low internet penetration rate and means that few Cubans have access to the content published by these blogs. While the same is true of the digital editions of national newspapers, evidence of significant ideological variance between the print editions available to Cubans and the content published online remains elusive.

In sum, the Cuba and the United States represent divergent models of mass media and communications. The American model is defined by its comparative freedom and association with the notion of democratic governance, as media organizations enjoy constitutional protection from undue state intervention. Nonetheless, publications like the *New York Times* face other obstacles, particularly an increasingly secretive state and growing economic pressures. By contrast, the Cuban model can be characterized as highly regulated and homogenous, with very little freedom afforded to those journalists who work within traditional news media organizations. Publications like *Granma* are directly connected to particular organs of the Cuban state, and can therefore be viewed as representative of the state’s positions and values on the issues they cover.

### 3.3 Data Sources

To identify framing patterns in coverage of Obama’s visit to Cuba, this project relies on a sample of 192 news articles from 90 publications. The majority of these data are obtained from the LexisNexis database through a query of newspaper articles containing the keywords ‘Cuba’ or ‘Cuban’ within 5 words of ‘United States’ or ‘American’, within 5 words of ‘visit' between March 1, 2016 and April 30, 2016. These criteria were chosen based on the premise that articles that mentioned these specified keywords in close succession were likely to be related to President Obama’s visit to the island, which occurred roughly halfway between the selected dates. Previous queries with wider search parameters tended to return a large number of articles that mentioned the event in question only superficially—for example ‘daily briefings’ that included one or two sentences related to U.S.-Cuba relations, or coverage of the 2016 American
presidential election wherein candidates provided a statement related to the normalization process. The publications represented include newspapers from the United States, Cuba, Canada, United Kingdom, India, Northern Ireland, China, Russia, Kenya, Nigeria, Australia, Qatar, Sri Lanka, South Africa, Iran, Singapore, Pakistan, and the United Arab Emirates. These articles serve two primary functions: to provide baseline context for the analysis of Granma and the New York Times, and to assess the importance of corpus size in the application of culturomic techniques.

Within these 192 articles, I take content from two daily newspapers as cases for study—Cuba’s Granma and the American New York Times. Though the latter was included within the LexisNexis query results, the former was obtained manually through transcription as no articles from the Cuban publication were available within the LexisNexis’ database. Each of the specific cases published a collection of articles grouped under the theme of Obama’s visit, though Granma’s content included articles published prior to the date of the visit which were deemed relevant to the event—this includes 47 articles from the Times and 12 from Granma, in addition to the remaining 133 articles from other publications.

While the differences in how these two newspapers operate are clear, they share a number of important characteristics. Both newspapers are published daily and sit near or at the top of their respective countries’ circulation rankings. Both papers publish digital editions and offer both English- and Spanish-language content. As discussed in the introductory section, there is also reason to believe that the two papers attract overlapping audiences, and represent divergent perspectives on foreign involvement in Cuban political and economic affairs. In order to control for as many causal factors as possible in determining frames through natural language processing, it was also crucial to address the issue of language itself in mining data for analysis. Studies examining the differences in framing across languages have suggested that representations of events and subject areas differ across linguistic and cultural contexts, and it was therefore important to ensure as much uniformity as possible in the language of publication for textual data (Correa, 2010). With this in mind, and given the focus on Anglophone readers outside of Cuba articulated in the introduction, the decision was made to focus on the English-
language content published by the *New York Times* and *Granma* in order to maintain consistency across all source materials.

While the examination of these two cases is substantively interesting, it also allows for various metrics observed in the two publications to be compared to a larger and more diverse dataset. With respect to the methodological question underpinning the study—whether ‘culturomic’ methods can be successfully applied on a smaller scale—the consideration of narrowly defined cases in comparison with a much larger sample was undertaken to assess the effect of corpus size on the results of each analytic component. It is important to acknowledge the difference between the size of this project’s content sample and what is common in other culturomic studies, as the analysis that is undertaken here does not involve sufficient data that the “high-throughput” computing Michel et al. (2011) specify in their definition of culturomics is required. The methodology presented here is designed to demonstrate the ways in which culturomic methods can be used to explore narrowly defined topics on a smaller scale than previous applications of computational techniques. This analysis involves little more than a few free software tools and a personal computer, and I have therefore refrained from labelling it as purely “statistical” or “quantitative” in nature. Instead, it provides an example of the insight that can be obtained through the application of computation to sets of data with well-defined methodological parameters.

### 3.4 Method

This project focuses on the headlines and body text of each article contained within the sample corpus, which is analyzed using a hybrid approach involving both automated computational techniques and qualitative interpretation. In general, the sourcing of data and interpretation of results rely heavily on the latter method, while the processing, storage, and analysis of data are accomplished through computation. The frame analysis undertaken in this study is inductive, in that it attempts to identify frames based on interpreted similarities among latent topics in a corpus rather than categorizing articles based on pre-defined frames. This approach is informed by two factors: first, the lack of scholarly attention paid to media framing of the U.S.-Cuba relationship has created a gap in understandings of the particular qualities of this subject area. Second, the
techniques employed to identify frames lend themselves to the exploration of unstructured content like mined news articles, particularly topic modelling and semantic analysis.

To answer the substantive and methodological questions proposed, this project’s analysis employs several natural language processing (NLP) techniques, referring to the computational analysis of natural language. In this context, natural language refers to language used in human communication, rather than artificial languages used in programming or mathematical notation (Bird et al., 2009). Natural language data are processed with the help of the programming language Python, which has become increasingly popular as a flexible, efficient means for data analysis. Like other programming languages, Python can receive a pre-written script and perform a particular function or set of functions based on the contents of that script. It can also make use of pre-written scripts within ‘toolkits’ or ‘packages,’ which are generally made freely available by programmers, research institutions, and academics across subject areas. In this study, a combination of original and pre-written scripts are used to store, process, and analyze data.

Prior to analysis, it was necessary to process the raw data that was collected from LexisNexis, though the three analytic components required different processing protocols. To identify and contextualize ngrams in the corpora under examination, raw textual data was tokenized. This refers to a process wherein sequences of text are broken up into individual words and symbols, known as tokens. From this set of tokens, those that were not useful or relevant to this study’s goals were removed—particularly stop words (i.e. insignificant words that are very common in natural language), punctuation, and other insignificant characters that appeared in the raw data. The removal of stop words is particularly important, as common words like “the,” “as,” or “and” would inevitably be the most common in any corpus and would therefore obscure more substantively significant aspects of the content. It became clear after initial applications of ngram analysis that other irregularities in the text would need to be addressed before meaningful results could be obtained. For example, the prevalence of various Spanish-language terms and phrases in the data led to typos in the English-language coverage—Raúl Castro’s name was often misspelled as ‘Raul Castro’, for example. After correcting
these errors and removing the aforementioned insignificant features, ngram frequency was assessed using a Python script.

Here, the frequencies of unigrams (one-word sequences), bigrams (two-word sequences), and trigrams (three-word sequences) are examined in relation to the overall size of the corpus under consideration. Rather than considering the absolute or raw frequency of ngrams (i.e. the number of times that ngram appears in the corpus), the comparison of frequencies across corpora requires that these frequencies be normalized. This simply refers to the division of each ngram’s total number of occurrences by the total size of the corpus, after which point the figure is multiplied by 10,000 to produce a measure of that ngram’s expected occurrences per 10,000 words. The formula is as follows:

$\text{Normalized Frequency} = \frac{\text{Raw Frequency}}{\text{Total number of ngrams in Corpus}} \times 10000$

For this calculation, articles published by the New York Times and Granma were separated from the remaining articles in the database to create three corpora, so that their frequencies could be considered individually and would not affect measurements in the other corpora. Ngrams for each corpus were first ranked based on their normalized frequencies, after which point a table was created for the unigrams, bigrams, and trigrams that were shared between the three corpora. This method allowed for the examination of each sample’s data individually and based on common features, though the frequencies of shared ngrams were often drastically different. Python scripts were written to find occurrences of a particular search term in each table, as well as to identify which terms were present in one set of data but absent from the others.

To create topic models, the corpora were separated once again but their data were processed differently. Where ngram frequency provides a direct measure of a term’s salience within and across corpora, the type of salience measured through topic modelling is more complex. The process involves the identification of ‘topics’ across a set of documents in a corpus, and can be approached using several techniques. Many involve the use of a TF-IDF matrix, an acronym for term frequency-inverse document frequency, wherein terms in a corpus are assigned importance based on both the number of times they appear in each document and their frequency across the entire corpus.
However, the use of TF-IDF matrices rests on the assumption that all words in a corpus are significant as individual components of a corpus—often referred to as the ‘bag-of-words’ model—but are not considered in terms of their connection to each other. Instead, this study uses latent Dirichlet allocation (LDA) to model topics in the corpora, which produces two primary outputs, word distribution per topic and topic distribution per document (Kim et al., 2019). In other words, LDA identifies latent topics that are spread across each corpus, with topics represented by collections of keywords that are determined to be related through their distribution. This technique is applied using a Python package called LDAvis, which allows the researcher to specify the number of topics and keywords returned by the algorithm (Sievert and Shirley, 2014). The interface also provides a useful visualization of the results, which includes a map depicting the relative ‘distance’ between topics in a corpus, as well as a histogram depicting the prevalence of the keywords present in each topic.

The topic models produced by LDA provide the basis upon which evidence of frames is identified in this study, as they present collections of keywords that are deemed to have some degree of substantive similarity. Categories are identified based on the qualitative interpretation of the three sets of topic models, according to the following criteria: first, they must be present in at least 2 of the 3 corpora under consideration; second, they must be represented in at least one topic within those corpora’s LDA results; third, the topics in which they are represented must contain at least three keywords with a strong and reasonable association to the subjects described by that category. Within these categories, prospective frames are inferred through the connection of salient keywords with relevant contextual characteristics of Obama’s visit.

Finally, the semantic characteristics of the three corpora’s coverage were examined through the measurement of sentiment polarity and subjectivity at the level of documents. These measurements were assessed using a Python package called TextBlob, which provides pre-written scripts to calculate the relative levels of each metric in a defined passage of text. Broadly speaking, this process involves the referencing of a lexicon of words with associated sentiment polarity and subjectivity information, whose scores change based on the presence of various ‘modifier’ words that may affect its semantic significance. The sentiment polarity metric is represented by a numerical value
between -1 and 1, measuring the relative positivity, negativity or neutrality of a particular set of words or phrases. Subjectivity is measured in a similar manner, and assigned a score between 0 and 1. To aid in the interpretation of these results, each corpus’ scores were sorted into four bins representing the four quartiles that the numerical values could occupy. These bins were then labelled to describe their significance—for example, sentiment polarity scores between -1 and -0.5 were labelled ‘Very Negative,’ whereas scores between 0.5 and 1 were labelled ‘Very Positive’. This process was repeated for each corpus’ subjectivity scores, with scores from 0 to 0.25 representing the least subjective (or most objective) and 0.75 to 1 representing the most subjective. The distribution of scores for each metric was subsequently plotted in two separate histograms for each of the corpora, which are presented in Section 4.

The resulting information around ngram frequency, topic models, and semantic characteristics of the corpora provided visual and numerical data for discussion, which are interpreted based on the theoretical models of framing presented in Section 2. The first two methods are selected for their ability to assess salience to different degrees of complexity, and the third is included to offer context that may aid in the interpretation of that salience. All three are used to evaluate the relative utility of computational methods in examining a specific case of media coverage around a particular event, particularly through comparative analysis of multiple sources.

4. Analysis

This project’s analysis was designed to identify characteristic similarities across the three corpora assembled from news coverage, particularly the salience of terms and phrases that can provide insight into frames present in each corpus. As stated by Entman (1993), the process of framing necessarily involves selection and salience. In the absence of empirical information around the drafting and publication of the news content under examination, I argue that the identification and interpretation of salience in a corpus provides the most direct proxy for understanding this phenomenon through quantitative study.

4.1 Ngram Frequency Analysis
This study’s analysis of ngram frequency engages with the corpora in question through the examination of unigrams, bigrams, and trigrams. This approach follows the notion that while the frequency of unigrams can provide a broad portrait of the substantive qualities of the corpora, the analysis of the most common bigrams and trigrams in each corpus adds depth to the results of the former calculation. Examining the most common 2 and 3 word combinations provides context, either reinforcing or minimizing the significance of the usage frequency statistics described in unigram results. The sample collected contains a total of roughly 115000 words after removing stop words and punctuation, of which ~6600 come from the *New York Times* and ~5400 come from *Granma*. Across these three distinct corpora, there are 762 shared unigrams, 130 shared bigrams, and 27 shared trigrams.

There are several notable features of ngram analysis results, including insights that differentiate patterns in word usage between the general sample and the two specific cases. First, there seems to be a significant difference in the amount of coverage dedicated to each country’s political leader. This discrepancy first emerges in the results of unigram frequency analysis, where ‘obama’ was among the most common in each of the three corpora, with very slight differences in frequency between the samples. ‘castro’, by contrast, appeared less than half as often than ‘obama’ in the general sample and in the *Times*. Even more significant is the difference in the frequency of each term within *Granma*’s coverage, wherein ‘obama’ appears ~260 times per 10000 words and ‘castro’ appears only ~41 times.

*Figure 1 - Normalized frequencies of ‘obama’ and ‘castro’ across corpora*

<table>
<thead>
<tr>
<th>Ngram</th>
<th><em>New York Times</em></th>
<th><em>Granma</em></th>
<th>General Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘obama’</td>
<td>230.718523</td>
<td>259.836674</td>
<td>267.716996</td>
</tr>
<tr>
<td>‘castro’</td>
<td>93.935399</td>
<td>40.831477</td>
<td>94.861121</td>
</tr>
</tbody>
</table>

However, the unigrams ‘obama’ and ‘castro’ lack the necessary context to affirm the hypothesis that the American president was a more salient figure in each sample’s coverage. In fact, the examination of bigrams across the corpora reveals that ‘obama’ is used to refer to members of the president’s family on multiple occasions, and ‘castro’
occurs when the erstwhile Cuban President’s late brother Fidel is mentioned, as well as Fidel’s late son Antonio. To minimize the influence of these other occurrences on our examination of Presidents Obama and Castro, three common bigrams that are commonly used to identify the two leaders are considered in parallel, with their frequencies added together to approximate total mentions of each individual. Though each of the three variations appear within the general sample of coverage, one is absent from each of the two specific cases.

As shown in figures 2 and 3, the examination of variations on the name of each leader lends further support for the notion that President Obama was mentioned more frequently than President Castro across the corpora. Both the general sample and the sample from Granma mention Obama roughly twice as much as Castro, and the Times mentions the American President more than 3 times as often as his Cuban counterpart.

Outside of the samples’ mentions of Presidents Obama and Castro, there are several ngrams that offer clues as to the topics that are most salient across the corpora. One of the most prominent is the issue of political dissent, which features heavily in both
the general sample of coverage and in the New York Times. In the general sample ‘political prisoners’ appears ~7 times per 10000 words and is also present in the trigram results as ‘existence political prisoners’ (~1/10000) and ‘political prisoners criticizing’ (~0.9/10000). The general corpus contains 63 bigram variations and 162 trigram variations on the term ‘prisoner’. It contains 174 bigram variations and 361 trigram variations on the term ‘dissident’. This topic appears to be even more salient in the Times, as ‘political prisoners’ is the 5th most common bigram with ~21 occurrences per 10000 words and ‘list political prisoners’ appears ~8 times per 10000 words. In total, the Times corpus contains 16 bigram variations and 37 trigram variations on ‘prisoner’. The Times corpus also contains 30 bigram variations and 45 trigram variations on the term ‘dissident’. By contrast, neither ‘prisoner’ nor ‘dissident’ appear as unigrams, bigrams, or trigrams in Granma’s coverage.

While the absence of these terms could be interpreted as a gap in Granma’s coverage of issues around dissent and imprisonment, ngram analysis results suggest that the broader subject of ‘human rights’ is distributed in a very different manner. In fact, the frequency distribution of variations on ‘human rights’ shows only a marginal difference between the general sample and Granma, which actually contains the highest frequency of the three corpora. As shown in figure 4, the bigram appears ~26 times per 10000 words in Granma—slightly more frequently than in the general sample and more than 3 times as often as in the Times. The trigram ‘issue human rights’ is one of the 27 that appear in each of the three corpora.

![Figure 4 - Normalized frequencies for variations on ‘human rights’](image)

<table>
<thead>
<tr>
<th>Ngram</th>
<th>New York Times</th>
<th>Granma</th>
<th>General Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘issue human rights’</td>
<td>1.648533</td>
<td>1.856665</td>
<td>0.097495</td>
</tr>
</tbody>
</table>

Human rights, dissent, and political prisoners are not the only ngrams that are distributed unevenly across the corpora—the ngram analysis results also provide evidence of significant differences in the language used to describe U.S.-Cuba relations. ‘Blockade,’ for example—a term long used in Cuba to describe the embargo that was
placed on the island beginning in the late 1950s—has a frequency of ~48 in the *Granma* corpus, meaning it appears more often than the unigram ‘castro’. By contrast, it’s frequency is only ~3 in the general sample and it does not appear in any of the *Times* coverage. There is a similar imbalance in the use of ‘communist’, which appears between ~15 and ~16 times per 10000 words in the general sample and the *Times*, but is completely absent from *Granma*’s coverage. Figure 5 shows several ngrams that are present in the general sample, but appear in only one of *Granma* or the *Times*. Of these terms, those that are uniquely present in *Granma* tend to fall further from the frequencies in the general sample than those that are unique to the *Times*. In cases when words appeared multiple forms, their frequencies are added together.

*Figure 5 - Normalized frequencies for selected ngrams unique to NYT or Granma*

<table>
<thead>
<tr>
<th>Ngram</th>
<th>New York Times</th>
<th>Granma</th>
<th>General Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘prisoner(s)’</td>
<td>24.719841</td>
<td>—</td>
<td>8.286942</td>
</tr>
<tr>
<td>‘dissident(s)/dissent’</td>
<td>29.663810</td>
<td>—</td>
<td>27.395657</td>
</tr>
<tr>
<td>‘communist(s)/communism’</td>
<td>16.47989</td>
<td>—</td>
<td>15.20897</td>
</tr>
<tr>
<td>‘journalist(s)’</td>
<td>18.12788</td>
<td>—</td>
<td>7.40950</td>
</tr>
<tr>
<td>‘blockade’</td>
<td>—</td>
<td>48.25538</td>
<td>2.92480</td>
</tr>
<tr>
<td>‘citizen(s)’</td>
<td>—</td>
<td>27.83964</td>
<td>9.55435</td>
</tr>
<tr>
<td>‘progress’</td>
<td>—</td>
<td>22.27171</td>
<td>4.777179</td>
</tr>
<tr>
<td>‘bilateral(ly)’</td>
<td>—</td>
<td>14.84781</td>
<td>3.50976</td>
</tr>
</tbody>
</table>

While the majority of salient ngrams seem to make reference to political and diplomatic subjects, there is some evidence of engagement with cultural topics in the coverage as well. Perhaps the most visible thematic characteristic of the ngram analysis results is the ubiquity of baseball as a topic across the corpora. The topic is well-represented in the general sample of coverage, but is particularly salient in the *Times* and only slightly less salient in *Granma*. In general, references to baseball are perhaps the most common outside of terms associated with the politics, policy, logistics, and personnel associated with Obama’s visit to Cuba. ‘Baseball’ has the 39th highest frequency in the general sample, the 15th highest frequency in the *Times* corpus, and the 97th highest in *Granma*. 
In the Times, 4 of the 25 most frequently occurring bigrams are references to baseball, while the most common trigram in the corpus is ‘tampa bay rays’—a reference to the professional baseball team based in Tampa, Florida that played an exhibition game in Cuba during the visit. This is not the only baseball reference, with “major league baseball,” “exhibition baseball game”, and “national baseball team” all appearing in the 20 most common trigrams across the corpus. In Granma, ‘tampa bay rays’ is the 8th most common trigram in the corpus and the only reference to baseball in the 25 most common bigrams or trigrams. Figure 6 shows ngrams referencing baseball across all three corpora.

Figure 6 - Normalized frequencies for ngrams referencing baseball

<table>
<thead>
<tr>
<th>Ngram</th>
<th>New York Times</th>
<th>Granma</th>
<th>General Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘baseball’</td>
<td>39.551747</td>
<td>14.847810</td>
<td>23.885894</td>
</tr>
<tr>
<td>‘national team’</td>
<td>11.537828</td>
<td>3.712642</td>
<td>7.214585</td>
</tr>
<tr>
<td>'baseball game'</td>
<td>11.537828</td>
<td>9.281604</td>
<td>6.824608</td>
</tr>
<tr>
<td>‘tampa bay rays’</td>
<td>14.836795</td>
<td>9.283327</td>
<td>7.897123</td>
</tr>
<tr>
<td>‘major league baseball’</td>
<td>9.891197</td>
<td>1.856665</td>
<td>3.022356</td>
</tr>
</tbody>
</table>

4.2 Topic Modelling

In order to contextualize and supplement the information unearthed by ngram frequency analysis, this study draws on the results of topic modelling through Latent Dirichlet Allocation (LDA). This model was applied to each corpus using a package for Python called LDAvis, which returns a visualization depicting the relative similarity of each topic as well as the top 30 ‘keywords’ associated with each one. The application of LDA to each corpus results in 10 ‘topics,’ each of which is assigned a number between 1 and 10 that corresponds with the percentage of tokens from the corpus that fall within that topic. That is, Topic 1 in each set of results is the most widely represented in that corpus, with Topic 10 having the narrowest representation. For example, Topic 1 in the general sample’s LDA results is associated with 24.1% of the tokens in that corpus. In spite of the relatively small datasets that were interpreted, topic modelling produced strong evidence to suggest that coverage of Obama’s visit to Cuba included several
distinctly salient topics—though some contained more uniformity in their constituent keywords than others.

In general, topic modelling is used to identify abstract semantic structures across large sets of documents. Unlike ngram frequencies, topics do not represent the most frequently occurring sets of words or phrases in a particular document. Rather, "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words” (Blei et al., 2003, 996). The usefulness of LDA rests on the assumption that documents can contain multiple topics simultaneously, and multiple topics can share the same keywords. The result of LDA is therefore a set of topics that transcend the limits of individual documents, and are uniquely distributed across a given corpus. This paradigm suits the identification of multiple frames across many documents, and eliminates the need to consider documents individually based on the predominant frames they contain or are associated with.

In this analysis, I define frames in the three corpora through a multi-stage coding process that begins with the automated clustering of keywords and topics, and concludes with the qualitative interpretation of those keywords and topics. This portion of the analysis follows the same method applied to the ngram frequency analysis, wherein corpora for the New York Times and Granma are separated from the remaining articles in the sample so that the three can be compared with one another. Rather than seeking evidence of frames that have already been identified in academic literature as common in news media coverage, I elected to identify frames based on the presences of reasonably similar or relevant sets of keywords in LDA results that can be connected to contextual characteristics of the Obama visit or the U.S.-Cuba relationship. For example, I argue that topics containing keywords like “price”, “markets”, “service”, or “company” alongside “progress” and “growth” could be interpreted as evidence of commerce and trade being framed as a means for alleviating economic hardship in Cuba.

After identifying keywords with reasonable associations within each corpus’ LDA results, I determine categories’ presence based on the following conditions—first, they must be present in at least 2 of the 3 corpora under consideration. Second, they must be represented in at least one topic within those corpora’s LDA results. Third, the topics in which they are represented must contain at least three keywords with a reasonable
association to the subjects included within the category. In some cases—particularly in the *Times* results, wherein the highest proportion of shared keywords occurred and the fewest clearly-defined frames could be observed—keywords associated with a particular category that occurred widely within that corpus’ results were identified. Following this method, I identified the following categories across the LDA results: *communication and dialogue; politics, policy and governance; commerce and trade; travel and tourism; and culture and sport*. Figure 7 shows the topics within each corpus that correspond to each of these categories.

*Figure 7 – Category-topic associations for the NYT, Granma, and general sample*

<table>
<thead>
<tr>
<th></th>
<th><strong>New York Times</strong></th>
<th><strong>Granma</strong></th>
<th><strong>General Sample</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Communication and Dialogue</strong></td>
<td>Scattered keywords (&quot;news, said, meeting&quot;)</td>
<td>Topic 2 (&quot;notes, press, explains, said, forum, noted&quot;)</td>
<td>Topic 1 (&quot;says, speech, published, twitter, history, television&quot;), Topic 2 (&quot;says, published, press, meeting, internet, twitter, journalists&quot;)</td>
</tr>
<tr>
<td><strong>Politics, Policy, and Governance</strong></td>
<td>Topic 1 (&quot;Castro, Raúl, meeting, revolution, officials, country, government, relations&quot;), Topic 2 (&quot;Castro, government, revolution, relations, country, national, officials&quot;)</td>
<td>Topic 3 (&quot;minister, countries, embassies, normalization, sovereignty, bilateral&quot;), Topic 7 (&quot;army, councils, general, ministers, talks, governor&quot;)</td>
<td>Topic 4 (&quot;secretary, department, regime, process, security&quot;), Topic 5 (&quot;Washington, ties, regime, diplomatic, officials, presidential, dissidents, Rhodes, republican, Rubio, nations&quot;), Topic 6 (&quot;Guantanamo, base, attacks, naval, peace, federal&quot;)</td>
</tr>
<tr>
<td>Category</td>
<td>New York Times</td>
<td>Granma</td>
<td>General Sample</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------</td>
<td>--------</td>
<td>----------------</td>
</tr>
<tr>
<td><strong>Commerce and Trade</strong></td>
<td>Scattered keyword (&quot;business&quot;)</td>
<td><strong>Topic 1</strong> (&quot;business, collaboration, sector, commerce, blockade&quot;)</td>
<td>Topic 3 (&quot;company, companies, airlines, flights, percent, service, cruise, tourism, market, Tampa, passengers, international, remittances, million, transportation, airports, Miami, restrictions, served, traveling, license, approval, popular, number, restaurants, health, price, markets&quot;), <strong>Topic 9</strong> (&quot;accounts, banks, transactions, dollars, salary, trade, treasury, llc&quot;)</td>
</tr>
<tr>
<td><strong>Travel and Tourism</strong></td>
<td>—</td>
<td><strong>Topic 4</strong> (&quot;arrivals, visitors, tourism, tourist, growth, festival&quot;)</td>
<td>Topic 3 (&quot;flights, airlines, cruise, tourism, passengers, transportation, airports, traveling, restaurants&quot;), <strong>Topic 7</strong> (&quot;Starwood, hotels, &quot;airport&quot;)</td>
</tr>
</tbody>
</table>
| **Culture and Sport**    | Scattered keyword ("baseball") | **Topic 1** ("stadium, gran, teatro"), **Topic 5** ("baseball, theatre, estadio, team"), **Topic 8** ("league, game, baseball, team, rays, stadium"), **Topic 9** ("concert, rolling, stones, music, band, event, tour, British") |}

The communication and dialogue category includes information related to press coverage of the visit, speeches delivered, and social media dynamics around the event. LDA analysis of the general corpus resulted in the two common topics including several top 30 keywords associated with diplomatic communications, public discourse, social media discussion, and press coverage of the period around President Obama’s visit. Topic 1 in the general sample’s LDA results contains the keywords “says”; “speech”; “published”; “twitter”; “history”; and “television”. Topic 2 contains a very similar set of keywords, including “says”; “published”; “press”; “meeting”; “internet”; “twitter”; and
“journalists”. Evidence of the communication and dialogue frame was also present in one of the topics returned in Granma’s LDA results—Topic 2 contains “notes”; “press”; “explains”; “said”; “forum”; and “noted”. Evidence of this category was not as clearly grouped in the Times corpus results, though keywords including ”news,” “said,” and “meeting” were scattered across many of the topics.

The politics, policy, and governance category included topics with keywords that included the names of government and military officials, terms associated with governance (including diplomacy), references to the capital cities or possessed foreign territories, trade negotiations, and national security. In the general sample, Topic 4 included the terms “secretary”; “department”; “regime”; “process”; and “security”, while Topic 5 included “Washington”; “ties”; “regime”; “diplomatic”; “officials”; “presidential”; “dissidents”; “Rhodes”; “republican”; “Rubio”; “nations”. The names “Rhodes” and “Rubio” likely refer to the White House staffer and presidential speechwriter Ben Rhodes and the United States Senator from Florida Marco Rubio, both of whom worked directly on issues around U.S. relations with Cuba. Topic 6 seems to make reference to military affairs, containing keywords including “naval”; ”benefits”; “Guantánamo”; “base”; “law”; “status”; and “cost”. The Granma corpus returned two topics that contain keywords related to politics, policy and governance, including Topic 3 (“minister, countries, embassies, normalization, sovereignty, bilateral”) and Topic 7 (“army”, “councils”, “general”, “ministers”, “talks”, “governor”). Granma’s coverage contains a particularly high proportion of keywords that reference national institutions and political or military leaders—as opposed to the references to domestic politics, dogma, or regime type that appear in the general corpus.

The politics, policy, and governance category also manifests in two topics across the New York Times corpus, though the difference between the keywords present in each is negligible. Topic 1 contains “Castro”, “Raúl”, “meeting”, “revolution”, “officials”, “country”, “government”, and “relations”, while Topic 2 contains these keywords alongside “country”, “national”, and “officials”. For context, the Cuban state apparatus is occasionally referred to as the la Revolución (“the Revolution”) within Cuba, and there exist many “Committees for the Defense of the Revolution” across the island that are charged with localizing government initiatives and liaising with national political
institutions. With that in mind, it is possible that “revolution” could refer to Fidel Castro’s initial seizure of power in Cuba or the current national government.

The commerce and trade category includes keywords associated with firms, financial trends and services, regulation and licensing, market forces. Though there is certainly a great deal of overlapping between the keywords associated with commerce and trade topics and those labelled travel and tourism-related, the latter category was differentiated to offer more nuance. Commerce and trade was most strongly represented in Topic 3 of the general sample’s results, which contained more than 25 keywords that have an association with commercial activity and trade. This includes “company”, “companies”, “percent”, “service”, “market”, “remittances”, “restrictions”, “license”, “price”, and “markets”. The high concentration of relevant keywords around this particular topic represents some of the strongest evidence for the prevalence of a frame in the dataset, with over 80 percent of the topic falling within the boundaries for the commerce and trade frame. Topic 9 contains several keywords that seem to reference financial transactions, including “accounts”, “banks”, “transactions”, “dollars”, “salary”, “trade”, “treasury”, and “llc”. Together, these two topics account for 18.3% of all tokens present in the general corpus. Granma also produced results that suggest a strong focus on commerce and trade in its coverage, with Topic 1 containing the keywords “business”, “collaboration”, “sector”, “commerce”, and “blockade”. This is especially strong evidence for the presence of a commerce and trade category in Granma’s coverage, as this topic represents 37.4% of all tokens in the corpus. While the New York Times results do not contain a topic that is clearly associated with commerce and trade, the term “business” does appear across all ten of the topics included in the results to varying degrees of relevance.

While it is present in some of the topics that make up the commerce and trade category, travel and tourism refers to topics and terms that are more specifically related to mass transportation, tourism, accommodations and other leisure activities. In the general sample, Topic 3 contains terms including “flights”, “airlines”, “cruise”, “tourism”, “passengers”, “transportation”, “airports”, “traveling”, and “restaurants”. Topic 7 contains “starwood”, “hotels”, and “airport”, with the latter two terms referring to the American hotel firm that was granted access to the Cuban market around the time
of President Obama’s visit to the island. In the *Granma* corpus’ LDA results Topic 4 contained several terms with a strong association to hospitality and tourism, including “arrivals”, “visitors”, “tourism”, “tourist”, “growth”, and “festival”. This represented a significantly smaller percentage of tokens in the corpus, with only 8% associated with Topic 4. The *New York Times* corpus does not contain any topics with a strong connection to tourism and travel, nor do any keywords related to the topic appear in its LDA results.

Finally, the *culture and sport* category was included to capture topics outside of the other five categories, particularly those associated with sporting activity and the arts. In the general sample, Topic 8 contained a number of terms associated with baseball, including “league”, “game”, “baseball”, “team”, “rays”, “stadium”. To provide context, the term “rays” is likely to represent the Tampa Bay Rays of Major League Baseball, as “tampa” and “bay” also appear in the keywords associated with this topic. The Rays took part in an exhibition game against the Cuban national baseball team during Obama’s visit, which was attended by the American President and his Cuban counterpart, Raúl Castro. Topic 9 in the general sample seems to reference the Rolling Stones, who played a concert in Havana just after the President’s visit—it includes the terms “concert”, “rolling”, “stones”, “music”, “band”, “event”, “tour”, and “british”. In the *Granma* corpus, Topic 1 contains several terms connected to arts and culture, including “stadium”, “gran” and “teatro”—the latter two terms likely referencing the Gran Teatro de la Habana, a theatre located in the capital where President Obama delivered a speech during his visit. Topic 5 contained the terms “baseball”, “theatre”, “estadio”, and “team,” though they are not represented across much of the dataset with the topic making up only 3.3% of tokens. In the *New York Times* corpus, the term “baseball” appears across 8 of the 10 topics generated by the LDA analysis, and is most salient in Topic 7. However, like many of the previous categories there were no topics with a concentration of relevant keywords strong enough to serve as clear evidence of a frame in the *Times* coverage.

Beyond the topics and keywords generated through the topic modelling process, the LDAvis package allows for the visualization of topics based on their relative similarity to one another—termed *inter-topic distance*. This metric can be understood as a proxy for difference between topics, as it is calculated using the Jensen–Shannon
divergence method for measuring similarity between probability distributions. Of the three, the general sample contained the most diverse inter-topic distance map, with at least two topics present in three of the four plotted quadrants. Granma’s results also placed at least one topic in each of the four quadrants, with a larger distance between outlier topics than is present in the general sample results. The Times seems to contain the smallest levels of inter-topic distance, with all but one outlier falling in two of the four quadrants. Figure 8 displays the inter-topic distance maps for each corpus. In other words, the corpora containing content from the New York Times and Granma were far more homogenous in terms of the topics that could be identified, whereas the general sample contained a more diverse set of topics that were more clearly differentiated from one another than in the other two samples.

Figure 8 - Inter-topic distance maps for NYT, Granma, and the general sample

4.3 Analysis of Semantic Characteristics

The final analytic method applied to the corpora is semantic analysis, which includes the assessment of sentiment polarity and subjectivity at the level of individual documents. Similar to the ways in which ngrams offered useful insight into differences between the three samples of news coverage, analysis of the sentiment and subjectivity returned results that suggest important similarities and differences in each corpus. To obtain these data, an algorithm iterated over each article in the three corpora, analyzing the combined headline and body text from each. The resulting numerical figures
represent the average level of sentiment polarity and subjectivity within the document, which fall on a scale between -1 and 1 for sentiment, and 0 to 1 for subjectivity.

Sentiment polarity and subjectivity values for each corpus were sorted into 4 ‘bins’—one for each quartile of their respective range of scores. In relation to sentiment polarity, the lowest quartile (bin 1) represents very negative sentiment in the text, the next lowest (bin 2) represents slightly negative sentiment, bin 3 represents slightly positive sentiment, and bin 4 represents very positive sentiment. Subjectivity scores were divided into bins as well, with bin 1 representing the least subjective (or most objective) scores and bin 4 representing the most subjective scores.

Analysis of the general sample showed all articles falling within bins 2 and 3, the slightly negative and slightly positive categories. They were not evenly split, however, with the algorithm returning a majority of slightly positive scores with the remaining portion made up of slightly negative scores. Figure 9 shows the distribution of scores in the sample between the two bins.

Figure 9 - Sentiment polarity distribution of articles within the general sample

With respect to subjectivity, the general sample contained a large majority of articles whose average landed within bin 2, indicating that they were, on average, slightly subjective. The other three bins were represented in the corpus as well, with relatively
even distribution between bins 1 and 3. Bin 4, representing the highest levels of subjectivity, contained the smallest number of articles. Figure 10 shows the distribution of scores across the four bins.

*Figure 10 - Subjectivity distribution of articles within the general sample*

![Subjectivity distribution chart]

In the New York Times corpus, the distribution of sentiment polarity values was similar to the general sample—all averages landed within bins 2 (slightly negative sentiment) and 3 (slightly positive sentiment). Values within bin 3 were much more common, with 37 slightly positive scores compared to 9 slightly negative scores. In other words, ~80% of articles in the New York Times database contained slightly positive sentiment polarity, with the remaining ~20% within the slightly negative category. Figure 11 shows the distribution of scores between the two bins.
The distribution of subjectivity scores for *Times* articles was also similar to the general sample, though no articles fell within bin 4. Rather, a majority of scores indicated slight subjectivity, with the remainder split relatively evenly between bins 1 and 3. The proportion of scores outside of bin 2 is higher in the *Times* sample than it is in the general sample, but the difference is not significant. Figure 12 shows the distribution of scores across the three bins.

*Figure 12 - Subjectivity distribution of articles within the NYT sample*
Finally, the *Granma* corpus also saw all of its articles land within bins 2 and 3, though their scores were more evenly distributed between the two than in the other two samples. 7 of the 12 articles fell within the slightly positive category, with the remaining 5 being assessed as slightly negative. In other words, only ~58% of articles contained slightly positive sentiment polarity scores, with ~42% slightly negative—a much larger proportion of negative scores than was present in the *Times* corpus or general sample. Figure 13 shows the distribution of scores between the two bins.

*Figure 13 - Sentiment polarity distribution of articles within the Granma sample*

*Granma*’s subjectivity distribution was the most narrow of the three, in that it contained only scores within bins 1 and 2—indicating high objectivity and slight subjectivity respectively. Like the other two samples the score proportions were uneven, with 10 of the 12 scores in this sample being classified as slightly subjective and 2 classified as very objective. Figure 14 shows the distribution of scores between the two bins.
5. Discussion

Before interpreting the results of the study’s three analytic components, it is important to return to the research questions that guide this inquiry. First, what evidence regarding the frames present in coverage of Obama’s visit to the island can be extracted through the computational analysis of media content? Second, can computational methods associated with large-n studies be applied effectively on a smaller scale—specifically in the analysis of news coverage around a single event during a short period of time?

In addressing the first of these questions, the three modes of analysis were somewhat successful individually—they each produced information from which inferences about the presence of specific frames and framing practices can be made. Ngram frequency analysis allowed for the identification of salient terms whose connection to relevant topics can be inferred, comparison of particular terms’ salience across the corpora, as well as the identification of ngrams that were uniquely salient in one of the two specific cases under study. Topic modelling allowed for the identification of latent topics across all documents in the sample, which were then categorized based on
their keywords. The semantic analysis portion provided data describing the sentiment polarity and subjectivity of the sample at the level of documents, which contained several features that differ between the general sample and the specific cases.

In answering the project’s methodological research question—whether or not quantitative methods for content analysis are effective when applied to a relatively small set of data—the techniques selected for application offered differing degrees of utility. In addition to the substantive conclusions that are drawn from this analysis, the consideration of ngram frequencies also provides a useful methodological conclusion—particularly the necessity of combining computational techniques with traditional qualitative interpretation. This is visible in three aspects of the analysis: first, the differences between results of unigram and bi/trigram frequencies that emerge from examining the surnames of Presidents Obama and Castro; second, in the examination of ngrams present in the general sample that are highly salient in one case while being absent from another; third, in the process of identifying differences in the framing of issues through qualitative inquiry, which is based on the observation of meaningful results in the computational analysis of content. The other methodological insight is related to the effect of corpus size on topic models, as there was a clear difference between the definition of topics in the much larger general sample and the two other cases, *Granma* and the *New York Times*. Nonetheless, the results of each analytic component support the use of similar methodologies in exploratory research, as they were each successful enough to offer descriptive information which could be expanded upon in future research.

This section will present some of the notable aspects of each corpus demonstrated in the results of each analytic component, as well as any methodological considerations that arose, before presenting the qualitative inquiries regarding frames that were prompted by particular results.

### 5.1 Ngram frequency analysis

The analysis of ngram frequencies within the sample provides a sense of which topics are most salient, as well as what similarities and differences exist between the general frequencies and those associated with a particular case. In general, the *New York
Times contained ngrams whose frequencies were only marginally different to the broader sample, while Granma’s frequencies diverged slightly in specific areas.

First, the difference in normalized frequencies between ‘obama’ and ‘castro’ suggests that all coverage of the trip contained in this sample focused much more heavily on the American President than his Cuban counterpart. This is supported by the analysis of variations on both surnames, in which mentions of the American president occur between 42% and 68% more often than mentions of President Castro across the corpora. With respect to differences between the broad sample and the specific cases, the New York Times presents very similar frequency distributions to those present in the general sample and Granma appears to diverge most drastically. It is notable the extent to which ‘castro’ is under-represent in Granma’s coverage, as occurrences of the surname are almost 80% less common than in the other two corpora, and are almost 147% less common than ‘obama’ in the paper’s coverage. When considering bigram variations on each leader’s name, Castro is mentioned roughly 17% less in Granma than in the general sample, and 38% less in Granma than in the Times. Compared to mentions of Obama, variations on Castro’s name occur ~77% less in Granma’s coverage.

Overall, these data suggest that readers of content across the sample are presented with coverage that mentions Obama more than Castro, and that readers of Granma are presented with an even greater degree of emphasis on the American president given the relative lack of discussion around Raúl Castro. Beyond the revelation that President Obama was a more salient figure in the sample of news coverage, the fact that Cuba’s leader was mentioned significantly less in the state-run Granma than in the other samples of coverage is interesting considering the ‘cult of personality’ phenomenon that is commonly associated with mass media content under autocratic regimes. In other autocratic countries—China, for example, where President Xi Jinping has begun to display a tendency toward personalist rule (Luqiu, 2016)—the establishment of a cult of personality in state media is commonly associated with the consolidation of power. Though it can be difficult to precisely define what it is that constitutes a cult of personality in mass media, it stands to reason that a high degree of salience associated with the leader’s name would be an important proxy for identifying such a pattern. With this in mind, the under-representation of Castro in Granma suggests that the
organization’s behaviour follows a slightly different pattern. It is not clear why *Granma* would prefer that its readers be presented with fewer references to the Cuban president, but deeper examination of modifier words that commonly appear alongside each of the leaders’ names could offer a stronger basis for inference.

The examination of leaders’ surnames also has methodological implications, as the results of analyzing bigram variations are very different to those associated with only ‘obama’ or ‘castro’. When considering only the two surnames, both frequencies are highest in the general sample with ‘obama’ occurring least frequently in the *Times* coverage and ‘castro’ least frequently in *Granma*. This picture changes when the parameters are narrowed, as the consideration of bigram variations results in Obama being mentioned most often by the *Times* and least often in the general sample, though these differences are marginal. The difference in mentions of Castro’s name in *Granma* narrows as well, though he remains under-represented compared to the frequency of variations on ‘obama’. While this change in the frequency proportions does not necessarily contradict the results of analyzing the names individually, it does affirm the need to account for variables that may not be visible in simple unigram frequency measurements and that some degree of qualitative interpretation is crucial in obtaining meaningful results from frequency distributions.

Apart from the issue’s status as a historical point of dispute between the governments of Cuba and the United States, it is striking that both the general sample and the *Times* contain relatively high frequencies of variations on ‘political prisoners’ and ‘dissidents/dissent’ while they are completely absent from *Granma*’s coverage. These data are consistent with the Cuban regime’s long-held narrative that it does not arbitrarily imprison citizens, which President Castro reiterated during a joint press conference with President Obama on March 21st, 2016. Given the fact that this statement was made publicly and on the record during Obama’s visit—in combination with the observed salience of both ngrams across the sample—it is notable that neither of the aforementioned ngrams occur throughout *Granma*’s coverage. While it is impossible to state categorically whether or not an editorial decision to abstain from publishing those remarks was made by the paper, these results do provide strong circumstantial evidence.
By contrast, ‘human rights’ are discussed with almost the same degree of frequency in *Granma* as in the general sample of coverage, but the lack of references to dissent or prisoners suggests that they are being framed differently in the Cuban publication than in the rest of the sample. These data also align with the Cuban government’s repeated invoking of perceived American neglect for human rights in defense of its own conduct. On multiple occasions, including the same press conference where President Castro addressed the issue of political prisoners, the government of Cuba has attempted to differentiate their definition of human rights from the one proposed by the United States and others. When discussing the issue alongside Obama on March 21st, Castro alluded to the ongoing disagreement between the two countries regarding definitions of human rights by stating that he “find[s] it inconceivable that a government does not defend and insure the right to health care, education, Social Security with provision and development, equal pay and the rights of children” (Gorman, 2016).

The ngrams that appear frequently in *Granma* but are absent from the *Times* provide their own insights regarding differences in the salience of particular topics, and are equally insightful with respect to the Cuban paper’s treatment of American involvement in Cuba. The unigram ‘blockade,’ for example, has long been used in Cuba to characterize the embargo that the United States initiated in the late 1950s, and its high frequency in *Granma*’s coverage suggests that it continues to be an important aspect of the state’s narrative around the U.S.-Cuba relationship. This is significant given the true meaning of the term—rather than describing the imposition of commercial sanctions, it is more often used to refer to the physical separation of a region from another through military means. Though there has been no naval blockade imposed on Cuba since the early 1960s, use of the term persists on the island as has come to possess clear symbolic value in the regime’s narrative around U.S.-Cuba relations. The fact that the term is never used in the *Times* coverage and occurs quite infrequently in the general sample (~3 times per 10000 words) affirms the notion that it is at least somewhat unique to the Cuban lexicon.

The other ngrams that occur in *Granma* but not in the *Times* make allusions to the implications of the two countries’ thaw in relations, particularly ‘progress’ and ‘bilateral(ly)’. The former term also appears in 20 bigram variations within *Granma*’s
coverage, with a sum frequency of 44.55169 occurrences per 10000 words. Though the
general sample contains a greater number of bigram variations, they occur at the
significantly lower sum rate of 10.33440 occurrences per 10000 words. With these data
providing context, Granma’s coverage exhibits the highest statistical predisposition to
invoking ‘progress’ of the three samples. The same can be said of variations on the term
‘bilateral,’ of which 6 occur in Granma’s coverage at a sum frequency of 29.70113 per
10000 words, compared to 7.01959 in the general sample and none in the Times. The
salience of such terms suggests that readers of Granma are presented with a degree of
positivity or optimism regarding the future of the two countries’ emerging bilateral
relationship that appears to be unique to the publication, though it is certainly possible
that a similar message is conveyed by the Times using other language.

While there is a great deal to be gleaned from ngrams that are unique to one of the
two case samples, there is also significance in their shared features. One such feature is
the high salience of baseball across all three corpora—in fact, baseball is the most salient
cultural topic throughout the three samples, though the Times places particular emphasis
on the sport in its coverage with the unigram ‘baseball’ appearing almost 40 times per
10000 words. As a sport that enjoys high levels of popularity in both Cuba and the United
States, and given the fact that an exhibition game between an American Major League
Baseball team and the Cuban national team was attended by Presidents Obama and
Castro during the visit, it stands to reason that the topic would feature heavily in coverage
of the nations’ relationship. Nonetheless, as Turner (2010) points out, baseball has long
served as a pretext for diplomatic negotiations between the two nations, representing a
“cultural connection [that] has been unable to overcome decades of hostility”. The
prevailing of the topic suggests that this connection remains intact, as readers of each
content sample are presented with extensive discussion of the sport within the context of
U.S.-Cuba relations and Obama’s trip to the island.

5.2 Topic Modelling

In general, the topic models associated with each corpus are mixed in their
significance. While the analysis extracted many meaningful results that corroborate
details present in the ngram analysis, it was most useful when applied to the larger and
more diverse general sample. This is not particularly surprising, as the greater depth of this sample was always likely to produce richer results as in most statistical modes of analysis. Nonetheless, the fact that multiple shared categories could be identified in at least two of the corpora does suggest that such methods can produce useful—if somewhat broad—results.

The politics, policy, and governance category was perhaps the most common across the three datasets, with each corpus containing at least two topics that meet the criteria for inclusion. All three corpora contain topics that reference government and military officials, as well as various terms associated with diplomatic processes (e.g. ‘relations’, ‘country’/‘countries’, ‘officials’, ‘ministers’). Potentially significant differences emerge in comparing the general sample’s topics with those present in Granma, as the former contains references to ‘dissidents’ and the latter includes the unigram ‘bilateral’. Both of these terms emerge as uniquely associated with the Times and Granma respectively in the ngram frequency results, providing a clear discrepancy in how the two publications may frame discussion of political issues.

For example, the presence of ‘dissidents’ in Topic 5 of the general sample alongside terms like ‘diplomatic’, ‘nations’, and ‘ties’, among others, suggests that the subject of dissent may be closely associated with discussions of diplomatic relations between the two countries in that sample of coverage. The presence of ‘Rubio’ and ‘republican’ is also interesting, as it appears to connect the discussion of dissent with Florida Senator Marco Rubio. Rubio is the son of Cuban immigrants to the United States, and has opposed the normalization of U.S.-Cuba relations by citing the fact that Cuba’s government has a long history of arbitrarily imprisoning dissidents (Jordan, 2016). His name appears alongside ‘Rhodes’, likely a reference to President Obama’s former Deputy National Security Adviser, Ben Rhodes. Rhodes is widely reported to have been one of the United States’ lead negotiators during the lead-up to the thaw in U.S.-Cuba relations, and has frequently advocated for the pursuit of diplomatic cooperation between the two countries.

Evidence of the politics, policy, and governance category in Granma’s coverage can also be connected to the corpus’ ngram frequencies. Topic 3 in the paper’s LDA results invokes ‘bilateral’—a term that was highly salient in Granma’s coverage but
completely absent from the *Times* corpus—alongside the terms ‘normalization’ and ‘sovereignty’. This combination could be interpreted as the Cuban paper emphasizing the country’s desire to retain a high degree of agency in diplomatic negotiations, a perspective that is supported by a review of news reports and public comments by Cuban government officials. Concerns about protecting Cuban sovereignty during negotiations have been articulated by the nation’s government on several occasions during the periods before and after Obama’s visit. This also serves as clear evidence of the “anti-imperialist” theme in *Granma*, as the reader is presented with content that seems to emphasize Cuba’s right to self-determination in the face of powerful American political and business interests.

Overall, these results suggest that readers of content from the general sample are presented with content invoking dissent as a central political issue, while readers of *Granma* are presented with content emphasizing the importance of bilateralism between Cuba and the United States. In addition, the presence of this theme in both ngram frequencies and the topic models associated with each publication supports the notion that *Granma* and the foreign press continue to represent divergent perspectives on which issues are most important to the successful normalization of relations. The general sample’s emphasis of dissent suggests that the Cuban state’s concept of human rights continues to be a central topic in discussions of the country presented in foreign media, while *Granma*’s emphasis of bilateralism aligns with the country’s apparent belief that it should be able to have a say in what is acceptable within its borders. For readers of content from the two corpora, therefore, frames within the *politics, policy, and governance* category are used to present very different messages.

The *commerce and trade* category is also notable, in that it contextualizes the use of ‘blockade’ in *Granma*’s coverage and encompasses the largest number of relevant keywords across all topics in the general sample. Topic 1 in *Granma*’s results contains ‘blockade’ alongside ‘business’, ‘collaboration’, ‘sector’, and ‘commerce’. This combination of terms could be interpreted to mean that discussions of the ‘blockade’ in *Granma* occur largely within the context of international commerce as opposed to policy or travel. For readers of the publication, it would seem that the ‘blockade’ is presented as an impediment to economic growth rather than a social injustice, as has been the case
during other periods in the countries’ relationship. In the general sample, 28 of the 30 keywords within Topic 3 can be reasonably categorized within the commerce and trade category. This is the largest proportion of relevant keywords within any topic across the three sets of results, though some of those terms are interpreted to fall within the travel and tourism category as well. The general sample also contains a topic composed of terms related to financial transactions and institutions, which could be related to the subject of remittances—long-considered a crucial aspect of U.S.-Cuba relations with important implications for both Cubans and Cuban expatriates (Sullivan, 2018). Overall, the category’s high degree of representation in both Granma and the general sample suggests that commerce and trade is presented to readers as a centrally important theme in coverage of Obama’s visit.

The travel and tourism category is also well-represented in Granma and the general sample. Topic 4 in Granma’s results contains the unigrams ‘arrivals’, ‘visitors’, ‘tourism’, ‘tourist’, ‘festival’, and ‘growth’. The final term in that list can be interpreted as contextually important, as it suggests that the travel and tourism category is applied in discussions of tourism’s economic benefits. This association is also supported by the overlapping nature of the commerce and trade and travel and tourism categories. In the general sample, for example, Topic 3 contains a number of terms specifically related to mass transportation that merit the topic’s inclusion in the travel and tourism category while simultaneously containing a large number of terms that can be connected to the commerce and trade category. Topic 7 also overlaps into the commerce and trade category, as it references Starwood Hotels and Resorts Worldwide, the Marriott subsidiary that was granted a contract to operate 3 Cuban hotels just before Obama’s visit to the island. Starwood’s expansion into the Cuban market is a good example of the overlapping nature of these two concepts, as it has profound implications for Americans and other foreign groups looking to travel to Cuba as well as other large corporations that may see Starwood’s expansion as setting an important precedent. The fact of this overlap within the general sample suggests that readers of these articles are being presented with content that emphasizes the connection between commercial ties with Cuba and increased tourism opportunities, while readers of Granma are presented with content that connects tourism to economic growth for Cuba.
Finally, the *culture and sport* category connects all three sets of LDA results with the ngram frequency analysis results through the presence of ‘baseball’—Topic 5 in *Granma* and Topic 8 in the general sample directly reference the sport. Though there are no topics within the Times’ LDA results that contain a sufficient number of associated terms to merit inclusion in the *culture and sport* category, there is evidence of ambient salience within the corpus as the term “baseball” appears across 8 of the 10 topics. While the *culture and sport* category is not the most widely or deeply represented within the LDA results, it is one of 3 that can be directly connected to notable ngram frequency results alongside *politics, policy and governance* and *commerce and trade*. This connection suggests that these three categories are among the most salient for readers of content across the three corpora, and that the topics they represent have been elevated by the publications under consideration.

Beyond the keywords associated with each corpus’ topics, differences among their inter-topic distance maps are also methodologically significant as they highlight the limitations of working with smaller samples of data when carrying out topic modelling. The general sample contains the most varied distribution of topics among the three samples, which can be explained by the relative depth of the corpus compared to the two others. Both *Granma* and the *New York Times*, by contrast, showed similar inter-topic distributions—the majority of topics clustered in a relatively compact area, with one or two outliers and several very large or very small nodes. This pattern is particularly evident in the *Times* results, where all but one node fell within two quadrants and node size tended to be very large or very small. These features suggest that the majority of tokens fell within a small number of topics, and that there was a higher degree of similarity in the topics identified by LDA. This stands to reason, as a larger corpus drawn from a more diverse set of sources is likely to produce a more varied distribution than a single source.

To summarize, topic models for each corpus provide both evidence of salience among particular terms and topics associated with five categories: *communication and dialogue; politics, policy and governance; commerce and trade; travel and tourism; and culture and sport*. Results for three of these categories (*politics, policy and governance; commerce and trade; and culture and sport*) include either individual or multiple
keywords that are identified as significant in the ngram frequency results. Finally, the lack of clearly-defined topics in the Times results and the homogeneity in both Granma’s and the Times’ inter-topic distance maps, compared to the much larger general sample, suggest that corpus size has a powerful determinative effect on the number of significant topics that can be extracted.

5.3 Analysis of Semantic Characteristics

Given the fact that this study relied on document-level semantic analysis, it is difficult to make direct connections between the salience of terms and topics identified in the two previous sections and the semantic features of each corpus. The assessment of sentiment polarity and subjectivity at the level of documents tends to produce more conservative scores than would be likely at the level of sentences, as documents contain a higher number of modifiers affecting overall scores than sentences do. Additionally, these scores cannot be attributed to phrases that contain particular ngrams, making it difficult to infer semantic context for specific features of the corpora. Nonetheless, the distribution of sentiment polarity and subjectivity scores can be useful in making broad characterizations of each corpus’ content.

First, though most sentiment polarity scores were predictably close to neutral there was some degree of variance among the three. The distribution of scores between ‘slightly negative’ and ‘slightly positive’ sentiment was similar in the Times and the general sample, though the Times contained a slightly higher proportion of ‘slightly negative’ scores. These features suggest that readers of both samples’ coverage were presented with coverage that was largely positive at the level of documents. Granma’s scores showed a much more even distribution between the two categories with ~58% of articles assessed as slightly positive and ~42% slightly negative. This suggests that the paper’s coverage was more balanced than the other two samples in terms of overall sentiment polarity at the level of documents, and that readers were presented with a more even mixture of favourable and unfavourable content.

The balance of positive and negative sentiment in Granma’s coverage could be interpreted in a variety of ways, though it would be necessary to explore several variables in conceiving such an interpretation—particularly the specific ways in which the Cuban
state projects its influence over the publication, as well as the journalistic and editorial practices that produced the specific reporting in question. For example, it is possible that Granma’s state funding has enabled the publication to avoid the sensationalist style that has come to be associated with “shareable” content, whereas publications operating in a more competitive media environment (i.e. the New York Times and other publications present in the general sample) employ higher degrees of positive sentiment to encourage the distribution of their content through social media channels. Likewise, this higher proportion of negative sentiment in Granma could be interpreted as a means of managing public expectation surrounding the normalization of relations with the United States, or as a ploy to draw concessions from the American negotiators by presenting a skeptical portrait of the visit in national media.

Second, the subjectivity distributions for the corpora under consideration tended to be more varied, as scores fell into a minimum of three bins within both the Times and general sample. Those two samples also contained the only occurrences of ‘very subjective’ and ‘extremely subjective’ scores, as Granma’s articles all fell into the bins representing low levels of subjectivity. While this does suggest that the Times and general sample contained documents with levels of subjectivity beyond what was present in Granma, each corpus contained a majority of ‘slightly subjective’ scores and a minority of scores within the other bins. The final notable characteristic of the subjectivity results was Granma’s consistently low levels of subjectivity—each document’s average falls into the ‘very objective’ or ‘slightly subjective’ bins. Overall, readers of the Times and the general sample were presented with content that was mostly ‘slightly subjective’ but would occasionally display higher levels of subjectivity than were presented to Granma’s readers.

5.4 Detection of Frames in Coverage

Within the results associated with each analytic component—ngram frequency analysis, topic modelling, and semantic analysis—I identified two occasions on which there was sufficient evidence that a connection could be drawn through qualitative interpretation to frames in coverage. First, I interpreted the discrepancy in the application
of “human rights” between Granma and the New York Times as evidence of a difference in the publications’ framing of the topic, which prompted a qualitative review of articles in which the relevant ngrams occur. This inquiry produced strong evidence of human rights issues being framed in very different ways by each paper. In Granma, for example, the topic of human rights was invoked to praise Cuba’s social service provision, stating that “Cubans enjoy guaranteed human rights, such as universal access to health and education”.\(^1\) This mirrors the statement that President Castro made during the joint press conference with President Obama, and suggests that Granma has elected to follow the Cuban leader’s example in criticizing American shortcomings in this area to avert attention from his own country’s domestic issues. Human rights are also connected to the ongoing embargo on Cuba, as Granma paraphrases a religious leader’s suggestion that “the U.S. blockade violates the human rights of the Cuban people.”\(^2\) Additionally, the publication discusses the American and Cuban governments’ mutually expressed desire to resolve differences in their respective views on the subject of human rights\(^3\), as well as Cuba’s commitment to meet the “minimum standards” for combatting human trafficking, which earned the country praise from the American Under Secretaries of State for Civilian Security, Democracy, and Human Rights, Sarah Sewall.\(^4\)

By contrast, the New York Times repeatedly connects the notion of human rights with the Cuban state’s policies regarding dissent and arbitrary imprisonment. The publication places emphasis on the activists who met with President Obama during the visit\(^5\), and documents the online response to President Castro’s suggestion that Cuba does

\(^{1}\) “Minute by minute: Follow Barack Obama's visit to Cuba here”, Granma, March 20, 2016.

\(^{2}\) Ibid.

\(^{3}\) Ibid.

\(^{4}\) “U.S. withdraws Cuba from its unilateral list of traffickers in persons”, Granma, July 28, 2015.

not have any political prisoners by quoting content from the Cuban blog 14 y Medio, Miami’s El Nuevo Herald, and Jorge Ramos of Univision—all of whom posted extensive lists of dissidents imprisoned by the Cuban state. The publication also describes President Castro’s response to questions about Cuba’s human rights record at the joint press conference as “apprehensive,” as well as “defensive and argumentative”.

These differences are significant, as they denote radically different framings of human rights by Granma and the New York Times. The former publication framed the United States as a hypocritical imperial power, who criticizes other nations’ treatment of citizens while neglecting to offer fundamental social services to their own citizens and violating Cubans’ human rights themselves through the embargo. It also applies the notion of human rights to frame Cuba as a leader in the provision of particular services for its citizens, and heralds its own achievements in the area of human trafficking despite its own admission that the assessment being referenced “is developed unilaterally, and in a questionable fashion, by the United States”. The latter publication framed the Cuban state as a perpetrator of human rights abuses through their imprisonment of activists and crackdowns on public dissent, as well as their leader’s refusal to publicly acknowledge these abuses despite strong public pressure. The Times characterization of Castro as ““apprehensive,” “defensive and argumentative” therefore stands in direct contrast with the image presented in Granma, which seems to lavish praise on the role of Cuba’s leadership in defending the rights of Cuban citizens.

Second, the presence of the keywords “normalization”, “sovereignty”, and “bilateral” within the politics, policy, and governance LDA category and the ngram frequency results suggests that Granma places far more emphasis on these topics than the Times or the general sample. This discrepancy also presents an opportunity to examine the ways in which sovereignty manifests in Granma and the Times. In the former publication, the pursuit and defense of sovereignty is presented as a noble principle that is

---

7 Ibid.
8 “Obama Says Embargo Would End; Castro Denies There Are Political Prisoners”.
9 “U.S. withdraws Cuba from its unilateral list of traffickers in persons”.
deeply engrained in Cuban society. For example, *Granma* presents the American occupation of the Guantánamo Bay naval base as a violation of Cuban sovereignty, and published a piece describing how a group of young activists staged a protest called the “Festival of Political Song” that called for the “illegally occupied” territory to be returned to Cuba. In a minute-by-minute report of Obama’s visit, *Granma* describes the Cuban people’s determination to “defend” their sovereignty, and noted that President Obama’s expressed his respect for that sovereignty. The fact that this term is usually discussed in relation to the fortitude of Cubans or the imperialist tendencies of the United States affirms the notion that *Granma* applied the concept of sovereignty to frame its apprehension at American involvement in domestic affairs as a protective posture, adopted to defend the rights of its citizens.

In the *New York Times*, sovereignty is only mentioned on a few occasions—in reference to the Cuban continuing disinterest in democratizing, where Cuba’s preoccupation with its own self-determination is presented as a major obstacle preventing such a shift, and within a profile of the Cuban poet and national icon José Martí, wherein the principles of sovereignty that he articulated are suggested to have developed into an “obsession”. Unlike in *Granma*, where sovereignty is held up as a central societal principle that must be defended, the *Times* uses the concept of sovereignty to frame Cuba’s concerns as an irrational impediment to progress in the development of the two nations’ relationship. While this instance of framing is not represented in such equal measure between the corpora as the previous discussion of human rights, it does provide a clear example of the two publications under examination representing one topic in radically different ways—which are, at least partially, connected to the contexts in which they are operating.

---

10 “Minute by minute: Follow Barack Obama's visit to Cuba here”.
12 “Minute by minute: Follow Barack Obama's visit to Cuba here”.
6. Conclusions

As this study engages with two distinct lines of inquiry—one substantive, one methodological—the conclusions for each will be summarized separately, before offering some discussion of the project’s limitations and how future research might engage with similar questions. The three analytic components included in this study produced several insights regarding the way Cuban and American news coverage of President Obama’s visit to Cuba differed. More specifically, they illuminated various areas where the language used in Granma and the New York Times can serve as a basis for inference regarding the frames present in their coverage. Of the three components, the results of ngram frequency analysis provided the strongest evidence of significant differences in the salience of particular terms and topics across the corpora. Topic models served as a useful complement to some of this evidence, and semantic analysis provided information regarding the tonal characteristics of the three samples under consideration.

Ultimately, however, the identification of conclusive evidence of frames is limited by the size of the sample under consideration. While analysis of the 133 articles within the general sample provided richer and more insightful topic models, computational analysis of the smaller cases did not offer any clear benefit when compared to a conventional close reading of the content. These two corpora were small enough that they could easily be read and coded manually, and topic modelling failed to provide a strong basis for inference when applied. However, the combination of exploratory computational analysis with subsequent qualitative examination did prove to be a relatively efficient method for detecting prospective frames that could then be explored in more detail.

In answering this study’s two research questions—whether frames could be identified in coverage of Obama’s visit to Cuba using culturomic techniques, and whether these techniques are useful in analyzing relatively small samples of content—a new question emerged: why should one choose to employ computational techniques in the identification of frames, particularly when examining such a small sample? In other words, is there any advantage associated with computational analysis of this amount of content when compared to qualitative assessment? If the goal of one’s analysis is to obtain clearly defined frames through computation, I do not believe the techniques
associated with culturomics are particularly suitable when applied to small cases. The analysis of my corpora produced results from which inferences about frames could be made and supported through qualitative study, but beyond offering quantitative support for those inferences I do not believe that the application of these techniques produced knowledge that could not be obtained through close reading. For this reason, I believe that culturomics can be more impactful as a tool for detecting frames than defining them, particularly when applied to small cases. Nonetheless, the aforementioned inferences did lead me to identify two strong examples of differentiation in the framing practices of Granma and the New York Times.

6.1 Substantive Conclusions

First, the ngram analysis provided evidence of particular terms being used disproportionately within one publication’s coverage, which, in turn, led to the discovery of differences in the framing practices employed by both specific cases. Most notably, the high salience of “human rights” across the corpora, in combination with the lack of references to dissent or imprisonment in Granma led to a qualitative evaluation of these terms’ use in both specific cases’ corpora. This inquiry found that Granma connected the notion of human rights with the suffering of Cubans under the embargo, the country’s provision of universal healthcare and education, and the Cuban and American presidents’ mutually expressed desire to resolve differences in their respective views on the subject.15 By contrast, the Times mentions human rights almost exclusively alongside descriptions of the Cuban state’s policies regarding dissent and arbitrary imprisonment, elevating the profiles of various activists who had been imprisoned in its discussions.16 This suggests that both publications effectively framed the other nation’s government as a perpetrator

15 “Minute by minute: Follow Barack Obama's visit to Cuba here”, Granma, March 20, 2016.
of human rights abuses—the Americans through their imposition of the embargo and failure to provide universal social services in Granma, and the Cubans through their imprisonment of activists and crackdowns on public dissent in the New York Times.

Further, the fact that discussion of ‘political prisoners’ was particularly salient in the Times suggests that the publication dedicated an especially large proportion of its coverage to the topic—even among non-Cuban media sources. I interpreted this and other characteristics as support for the assertion that the New York Times projects various tenets of American liberalism in its coverage, which aligns with González’ (2017) perspective and differentiates the Times from Granma. Those terms that are unique to the Times are largely related to Cuba’s poor human rights record, and portrayals of human suffering as a justification for intervention have appeared in many media depictions of U.S. foreign policy, as discussed by Herman and Chomsky (1988) and Krishnaiah et al. (1993). However, as Mermin (1999) points out, this does not necessarily mean that the perspective advanced in news media originates with the source in question. Rather, it seems that “the spectrum of debate in Washington, instead, has determined the spectrum of debate in the news” (Mermin, 1999, 143). I believe that the evidence presented here supports Mermin’s assessment, as Cuba’s record of human rights abuses has long been cited as a barrier to the successful normalization of relations with the United States. The fact that the topic of human rights is prevalent in both historical public discourse and the sample of media discourse I analyzed suggests that the Times may be projecting the political establishment’s ideas to a significant extent in its coverage of Cuba.

Nonetheless, I do not observe sufficient evidence in my analysis of the Times to support González’ assertion that the New York Times presents such ideas as though they were “irrefutable truth” (185). While the manner in which the Times presents American liberalism in its coverage is not necessarily captured by the metrics I consider, I believe that the large proportion of articles that are assessed to be ‘slightly positive’ suggests that the publication’s coverage of the visit was largely favourable. If criticism of Cuban human rights abuses were as ubiquitous in the foreground of Times coverage as the author suggests, I believe that the semantic characteristics of the corpus would have appeared more balanced between positive and negative sentiment. Further, the salience of terms and topics related to imprisonment and dissent is certainly quite significant but the
fact that topics like baseball seem to be even more salient—at least in quantitative terms—undermines the idea that human rights issues occupy the most central position in the *Times’* framing of Obama’s visit.

Second, the presence of the terms “normalization”, “sovereignty”, and “bilateral” within the *politics, policy, and governance* LDA category suggested that *Granma* had emphasized the discussion of Cuban self-determination and bilateral cooperation during the process of normalizing relations more than the *Times* or the general sample. A close examination of *Granma*’s content supported this suggestion, as the publication emphasizes, at various times, the notion that the United States’ continued operation of the Guantánamo Bay naval base constitutes a violation of sovereignty\(^\text{17}\), the Cuban people’s determination to “defend” their sovereignty, and President Obama’s expression of respect for that sovereignty during the normalization process.\(^\text{18}\) These instances capture the sense of nobility associated with the defense of Cuban self-determination that is presented in *Granma*, where the term appears more than three times as often as it does in the *New York Times*. In the *Times*, sovereignty is only cited in reference to the Cuban state’s unwillingness to accept American democratic principles\(^\text{19}\), as well as a description of the country’s national hero, José Martí, wherein the paper discusses a general preoccupation with the topic among Cubans.\(^\text{20}\) I argue that the difference between these representations of sovereignty suggests that the term was applied to frame the normalization process in very different ways by each publication. *Granma* uses the notion of sovereignty to portray Cuba’s refusal to change its practices as a defense of national principles, and to discuss the concessions it hopes to obtain during negotiations through criticism of the American ‘imperialist’ occupation of Guantánamo. Indeed, the appearance of sovereignty alongside “progress” in *Granma*’s LDA results suggests that it is presented as crucial to the advancement of normalization. By contrast, the *Times* seems to apply sovereignty to

\(^{17}\) “Minute by minute: Follow Barack Obama's visit to Cuba here”.

\(^{18}\) Ibid.


frame Cuba’s unwillingness to make important concessions in a more pejorative fashion, as the term is presented as an “obsession” and an obstacle to democratization.

*Granma*’s application of sovereignty also provides an opportunity to connect the results of ngram frequency analysis and LDA with the semantic characteristics of its coverage. As the publication’s sentiment polarity scores show a somewhat even distribution of ‘slightly negative’ and ‘slightly positive’ sentiment in its articles, it stands to reason that *Granma* wished to emphasize the Cuban state’s reservations about normalization alongside a sense of relative optimism rather than attempting to seduce its readers with the historical significance and sense of romance that some associated with Obama’s visit. Likewise, these characteristics could be related to the ways in which ordinary Cubans were likely to interpret the visit and its significance, with the balanced sentiment presented by *Granma* representing the state’s desire to manage expectations around the future of Cuba. Overall, it is clear that the publication represents a significant divergence from the *New York Times*’ and the general sample’s framing of the visit and that its coverage places emphasis on topics that are less salient in the other two samples. If anything, these characteristics make it more difficult to disentangle the interests that are served by *Granma*’s coverage than a publication like the *Times*, despite its closeness to Cuba’s political class.

**6.2 Methodological Conclusions**

In answering this project’s methodological question—whether computational techniques for analysis can be applied effectively in the examination of relatively small and narrowly defined samples—the results were mixed. I believe that the usefulness of these techniques is equally dependent on the methodology employed as it is on the size of the sample under consideration, as a well-defined method can help to minimize the limitations associated with computational analysis of small samples.

While the examination of three separate corpora within the sample of articles had substantive utility, it was equally important to the methodological question as it provided the opportunity to compare the results of analyzing the smaller corpora with one that was much larger. In undertaking the ngram analysis, I argue that this method produced no significant difference between the results associated with each corpus beyond their total
size. As the general sample obtained from LexisNexis was significantly larger than the other two corpora, it contained a larger number of total ngrams and produced a larger lexicon of distinct terms. Nonetheless, the normalization of ngram frequencies allowed for useful comparisons based on shared terms, irrespective of the absolute size of the corpora. The comparison of the samples based on unique ngrams was also insightful, and was made particularly efficient through the use of Python to isolate and return each corpus’ unique terms.

The construction of topic models was more mixed in its utility, and was problematic when applied to the smaller corpora. While several well-defined topics were identified in the general sample, fewer could be drawn from the New York Times’ and Granma’s coverage. That the results of applying LDA to the general sample were so much more useful suggests that the size of the corpus under consideration has a profound effect on the results one can hope to obtain through topic modelling, and supports the notion that it is less helpful in asking questions of narrow samples. Nonetheless, it was possible to identify five content categories based on the interpretation of all three samples’ results, though the number and depth of categories would be likely to increase if the parameters of the data were to be expanded.

Finally, the analysis of semantic characteristics also offered mixed results in terms of its usefulness in studies of this nature. While it did help to offer interesting points of comparison between the three samples and aided in the contextualization of ngram frequencies and topic models, analyzing sentiment polarity and subjectivity at the level of documents provided very few opportunities to connect the salient terms and topics present in the other results to the semantic characteristics of the text. Nonetheless, I believe that the inclusion of this analytic component was particularly helpful as an exploratory tool. Its results provoked new questions that can be addressed through deeper study, as differences between the corpora confirmed that patterns in semantic characteristics seem to have a strong circumstantial connection to the type of publication under consideration and the event being discussed in coverage.

### 6.3 Limitations
The parameters of this study were limited by several variables. First, the computing resources and data that were available made it difficult to mount a particularly extensive computational analysis in this particular subject area. While it is fairly straightforward to obtain data from the *New York Times* and other sources present in the LexisNexis database, very few sources of Cuban media content were so accessible. There is also a lack of data regarding the digital media consumption habits of Cubans, which is understandable but made it particularly difficult to select a source for comparison with the *New York Times* and the rest of the general sample. Had I been able to identify which sources of digital news were most popular among Cubans—for example, ones obtained through the illicit distribution of media content for example, which is quite common on the island—I believe my comparison of cases could have been connected directly to readers in a way that was not possible otherwise.

Second, my own expertise proved to be a limiting factor with respect to the complexity of techniques that could be employed and the mining of raw data. Had I been more familiar with automated web crawlers, for example, it may have been possible to collect an even greater volume of data from Cuban media sources. Instead, I was limited to those that I could access manually and transcribe.

Third, the issue of language proved to be a significant limitation—both with respect to the logistics of the project and the substantive significance of its results. Though I may have preferred to analyze a sample that contained media content in the language that its readers were most likely to speak, I knew that it would be problematic to mount a comparison across linguistic contexts. It was also difficult to account for these differences, and the prospect of translating one or more samples between English and Spanish would have been too time intensive.

Nonetheless, there are limiting aspects of my method whose alteration could lead to an increase in the utility of this approach. One example is the expansion of the semantic analysis component to engage particular topics with more precision than is possible at the level of documents. Had it been possible to explore sentence-level sentiment polarity and subjectivity for occurrences of the categories, topics and keywords that were returned by LDA, the picture of my sample’s semantic characteristics would have been far more nuanced and, I suspect, informative. I believe that a more granular
portrayal of framing across the corpora would provide stronger conclusions regarding the manner in which particular topics are framed by the publications under consideration.

6.4 Future Research

In spite of the limitations specified above, I argue that the use of relatively small datasets can offer a useful perspective on the way computational methodologies function, and can affirm the utility of these methods in analyzing larger and more complex sets of data as it becomes available. In the future, similar methods may be applied to larger and more diverse samples of news coverage to examine other instances of media framing around a particular event across cultural, political, and/or economic contexts. It would also be useful in an epistemological sense, as the direct comparison of a conventional method for coding frames in news coverage with a computational approach would likely offer constructive feedback for both schools of thought around content and discourse analysis. This type of comparison would present an opportunity to identify empirical information that was lost or gained depending on the methodology employed, and to make recommendations for how hybrid approaches’ efficacy can be maximized.

With respect to frames in the Cuban media, several of this study’s substantive results merit closer examination. First, a wider analysis of Granma and other state-affiliated publications could provide insight as to how the messages presented by the Cuban state through its media partners have evolved over time. Though this study does not engage with the evolution of frames over a significant length of time, many other studies have demonstrated the usefulness of this variable in assessing large-scale trends in various social, cultural and political phenomena. This would be particularly interesting to observe as Cuba experiences a number of broad societal shifts in in the post-Castro era. The methodology presented here could also be replicated in a comparative analysis of different media actors within Cuba, particularly between state-run publications and the country’s vibrant blogging community. Such a comparison could provide valuable insight into differences in values and perspectives between the country’s political leadership and its citizens, and could help researchers predict which issues will be most salient in public discourse over the near- and medium-term. For anyone with an interest in Cuba, this
insight would be a great tool for understanding the nuances of a particularly complex and dynamic society.
References


Sean Gorman. “Obama’s Reply to Castro’s Criticism Open to Interpretation.”


