Quantitative Text Analysis with Asian Languages: Some Problems and Solutions

Amy Catalinac
Harvard University
acatalinac@fas.harvard.edu

Consider the following problem. We want to know how Japanese politicians feel about China. Perhaps we want to advise Australian government officials about how to respond to Japanese overtures for security cooperation or make a theoretical point about how power shifts impact enduring rivalries. We could interview or survey politicians, but this would be difficult. Those who are more likely to speak to a foreign researcher or seek out an opportunity to comment on foreign policy might be different from the average politician. Ensuring that our sample was not biased would consume a lot of our time and probably much of our social capital. It would yield estimates of nothing more than how these politicians felt about China on the day they did the interview or filled out the questionnaire. It might be difficult to compare their responses. It also requires the rather large assumption that politicians truthfully answer the questions put to them by researchers.

Methods for quantitative text analysis offer a solution. Pioneered by computer scientists, these methods enable the systematic analysis of large collections of texts in ways that allow us to draw valid inferences about the world in a timely, replicable fashion. They are being applied by political scientists to estimate the feelings, priorities, concerns, threat perceptions, policy preferences, ideologies, and other characteristics of political actors from the words they write or utter. These methods have enabled the drawing of inferences about Japanese politicians from their election manifestos [2], United States Senators from their press releases [5], Muslim clerics from their fatwas [12], the United States government from its treaties [15], Russian civil and military elites from their public statements [16], and the American public from their blogs [7]. Other applications have drawn inferences about actors from the way they treat the words of other actors. We can now say with greater confidence than ever before that government ministers became more accountable to opposition politicians between 1880 and 1900 in the British House of Commons [3], the Chinese government censors social media posts inciting collective action [9], and the deaths of prominent Jihadi clerics did not increase readership of their online material [11].

In a typical project using quantitative text analysis, the work flow is as follows:

1. **Conceive of Research Question and Quantity of Interest**: Identify a research question or relationship of concern, and conceive of the quantity of interest that would enable you to examine this relationship. To continue the above example, we might want to know whether concern about China among Japanese politicians fluctuates with the performance of the Japanese economy.
2. **Acquire a Corpus**: Obtain a collection of texts, called a “corpus", from which your quantity of interest can be estimated. We might choose the universe of parliamentary speeches uttered by Japanese politicians during a period in which both the security threat posed by China and Japan’s economic performance varied.
3. **Convert Corpus into Data**: Convert the corpus into quantitative data. In most instances this is achieved by constructing a “term-document matrix" from the corpus. This is a matrix that has document identifiers in the columns (“speech 1”, “speech 2”, “speech 3”, etc.) and the unique words that appear in the corpus in the rows (“word 1”, “word 2”, “word 3”, etc.). The cells denote the frequency with which each unique word appears in each document. The size of this matrix can be large and unwieldy, but can be pruned down before the application of the model.
4. **Select a Model to Estimate Quantity of Interest**: Select a model to estimate your quantity of interest. If we were interested in measuring concern about China from these speeches, we could proceed in several ways. We could choose to run a “topic model" on the speeches, which is a method that uses the frequencies with which words appear in the documents to uncover the latent topics from which those words were generated [13]. Topic models cannot be instructed to recover particular topics, however so we would have to hope that it uncovered a topic resembling concern about China. Or we could read some of the speeches, decide which ones indicated more or less concern about China, hand-code a set of the speeches ourselves, and teach a computer to code the rest [7].
5. **Use Quantity of Interest to Examine Relationship**: Use the estimates of your quantity of interest derived from (4) to examine the relationship you are interested in. For example, we might examine how our estimates of the level of concern about China varied with conventional indicators of economic performance such as unemployment.

Scholars interested in making inferences about countries where English is not spoken need to work with texts written in the native language. Working with texts written in Asian languages usually entails problems at the second (“acquiring the corpus") and third (“converting it into data") stages. This is because quantitative text analysis requires machine-readable text. Scholars can either choose research questions that can be answered with text that is already machine-readable, such as text located on the Web, or, if they choose research questions that cannot be answered with text already in this format, they can use optical character recognition (OCR) software to extract text from scanned images [4, 15], and save this text in a machine-readable format. Unfortunately, when languages have complicated scriptive representations or customs such as writing different sections of a page in different ways (from the top left, for example, or from the top right), OCR software tends to perform badly.

In my work with thousands of Japanese texts ranging from one to twenty five years old, I found that even native OCR software such as Yomitori Kakumei, Yonde Koko, and the Abbyy Fine Reader proved utterly incapable of extracting words from scanned images of the texts, even when those scanned images were of high quality. This is likely to remain a major impediment to the application of these methods to text written in complicated languages, that is not already machine-readable. Until better OCR software becomes available, my reluctant recommendation is for researchers working with these texts to choose research questions that can be answered with machine-readable text. While permission and access often need to be arranged, the digitization of the universe of debates in the Japanese Diet (available at http://kokkai.ndl.go.jp/), the universe of articles from Japanese newspapers in the postwar period and earlier (see https://database.yomiuri.co.jp/rekishikan/ for the *Yomiuri Shimbun* and http://database.asahi.com/library2e/ for the *Asahi Shimbun*), and other textual corpora written in Asian languages makes this plausible.

We could relax the requirement that OCR perform well if the errors it generated were random. But close-to-perfect extraction of text is a necessity for texts written in Asian languages for another reason: the absence of spaces between words. This means that scholars must use a “tokenizer" to parse out the text. Only once the text is parsed out can a term-document matrix be constructed. If the machine-readable text has not been extracted properly and contains errors, it follows that the tokenizer will fail to parse out the text, which will make the construction of a term-document matrix impossible.

Using a tokenizer can also confer advantages. The tokenizer MeCab (http://mecab.sourceforge.net/), for example, implemented in the R programming language by Motohiro Ishida [8] (http://rmecab.jp/wiki/index.php?RMeCab) not only parses out Japanese text, but also stems and categorizes each Japanese word. Stemming is a standard pre-processing step in natural language processing, while categorization gives the researcher the means of pruning the matrix to speed up the estimation. For example, when applying the tokenizer RMeCab to my Japanese-language texts, I was able to use the categories to remove words that performed purely grammatical functions, which were unlikely to convey useful information about the content of the texts. This pruning meant that the two models I applied to my Japanese-language texts, the “topic model" Latent Dirichlet Allocation [1] and the “scaling model" Wordfish [14], yielded intuitive estimates of the topics and ideological positions contained in the texts.

While the above two problems apply to all Asian languages, Japanese has another feature that makes quantitative text analysis difficult: heterogeneity in scriptive representations of the same word. Modern-day Japanese is comprised of three scripts: kanji, katakana, and hiragana. While all Japanese words can technically be written in hiragana or katakana, with the former being reserved for native Japanese words and the latter being used for phonetic translations of Western words, some can also be written in kanji. Some syllables within the words that can be written in kanji can also be written in hiragana. Unless there is a prescribed scriptive representation for each word, which could be the case if all the texts in the corpus were authored by a single person, the same words will appear in different scriptive representations within the corpus and sometimes even within individual texts. We cannot simply convert the universe of words in the corpus to a single scriptive representation because this would make words whose substantive meanings were different appear identical, which would then complicate interpretation of the model’s output.

Unfortunately, heterogeneity in scriptive representations of the same word poses difficulties for scholars interested in applying quantitative text analysis to Japanese-language texts written by different authors. It suggests that analysis of corpora authored by a single source, such as a government or newspaper, which might have prescribed scriptive representations for each word, will be easier. When such corpora are unavailable or unsuitable for examining the relationship of interest, two possible solutions are as follows. First, the researcher could construct a dictionary of words whose scriptive representations varied throughout the corpus and write a program to convert all scriptive representations to a single one. Alternatively, the researcher could apply the method she is interested in to a set of texts in which scriptive heterogeneity existed and the same set of texts after it had been removed. If the method produced identical or similar results across the two sets, she could justify using the uncorrected texts. Even if the results were biased in a particular way, she could still use the uncorrected texts and correct for the bias.

Notwithstanding these difficulties, scholars of Asia should use quantitative text analysis to answer questions they are interested in. I write this with some hesitation. Foreign researchers have typically invested years of their lives mastering an Asian language, only to now to told to turn their attention toward statistics. But while the costs of investing in new methodological skills are high, they are paid off quickly. They can also be lowered by seeking co-authors with complementary skill sets or applying for the funding to hire professionals adept in these methods. While it is unfathomable that a project using Asian languages could succeed without substantial area expertise, working with large textual corpora might enable scholars who are unable to constantly visit their countries of interest to continue researching it. It also means that those trips become about explaining to political actors and political scientists on the ground what has been observed in the texts and asking them why. In my experience, this leads to fruitful conversions that ignite future research.

For an excellent overview of the methods and applications in this area, see Justin Grimmer and Brandon M. Stewart’s “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts” [6]. For new applications in comparative politics, see “Computer assisted text analysis for comparative politics” by Christopher Lucas et al. [10]. Please note that the unpublished papers cited herein can be found on the Web.

References

[1] Blei, David M.and Andrew Y. Ng and Michael I. Jordan (2003), “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 3: 999-1022.

[2] Catalinac, Amy (2014), “Pork to Policy: The Rise of National Security in Elections in Japan”, *unpublished manuscript*.

[3] Eggers, Andrew C. and Arthur Spirling (forthcoming), “Ministerial Responsiveness in Westminster Systems. Institutional Choices and House of Commons Debate, 1832-1915”, *American Journal of Political Science*.

[4] Eggers, Andrew C. and Jens Hainmueller (2009), “MPs For Sale: Returns to Office in Post-War British Politics”, *American Political Science Review*, 103(4): 1-21.

[5] Grimmer, Justin,(2013), *Representational Style in Congress. What Legislators Say and Why It Matters*, Cambridge University Press.

[6] Grimmer, Justin and Brandon M. Stewart (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts" *American Political Science Review*, 21(3): 267-297.

[7] Hopkins, Daniel and Gary King (2010), "A Method for Automated Non-parametric Content Analysis for Social Science", *American Journal of Political Science*, 54(1): 229-247.

[8] Ishida, Motohiro (2010), *R ni yoru Tekisuto Mainingu Nyumon [Introduction to Text Mining]*, Morikita, 2010.

[9] King, Gary and Jennifer Pan and Margaret E. Roberts (2013), "How Censorship in China Allows Government Criticism but Silences Collective Expression" *American Political Science Review*, 107: 1-18.

[10] Lucas, Christopher, Richard Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer and Dustin Tingley (2014), "Computer assisted text analysis for comparative politics", *unpublished manuscript*.

[11] Nielson, Richard (2014), "Martyrdom or Irrelevance? The Effects of Targeting Jihadi Ideologues", *unpublished manuscript*.

[12] Nielson, Richard (2014), "Networks, Careers, and the Jihadi Radicalization of Muslim Clerics", *unpublished manuscript*.

[13] Quinn, Kevin M. and Burt L. Monroe and Michael Colaresi and Michael H. Crespin and Dragomir R. Radev (2010), "How To Analyze Political Attention With Minimal Assumptions And Costs", *American Journal of Political Science*, 54(1): 209-228.

[14] Slapin, Jonathan B. and Sven-Oliver Proksch (2008), "A Scaling Model for Estimating Time-Series Party Positions from Texts", *American Journal of Political Science*, 52(3): 705-722.

[15] Spirling, Arthur (2011), "U.S. Treatymaking with American Indians. Institutional Change and Relative Power, 1784-1911", *American Journal of Political Science*, 56(1): 84-97.

[16] Stewart, Brandon M. and Yuri M. Zhukov (2009), "Use of Force and Civil-Military Relations in Russia: an Automated Content Analysis", *Small Wars and Insurgencies*, 20(2): 319-343.