Philippine Social Sciences Review (2018) • Vol.70 No.1 • pp. 52-74 • © CSSP UP Diliman

John Taylor's *Philippine Insurrection against the United States*: A preliminary digital interrogation of the archivist

Nicholas Michael C. Sy

ABSTRACT

John Taylor's five-volume documentary collection *The Philippine Insurrection against the United States* is an invaluable resource on the Philippine-American War. However, the criteria with which Taylor selected the documents for his collection have remained opaque. The present methodological paper explores the use of Topic Modeling, a tool from the Digital Humanities, to posthumously interrogate the archivist. By comparing the preliminary results of successive manual and automatic coding, its findings highlight the crucial role that materials external to a corpus play in helping the researcher to gauge whether or not particular strings of co-occurrent words are relevant to a research project.

Keywords: Philippine-American War, Digital Humanities, Topic Modeling, MALLET, History

This research note documents and evaluates the author's application of Topic Modeling (TM), a tool in digital humanities, to a small sample of 107 out of 1,534 documents taken from an important published primary source collection in Philippine history: John R.M. Taylor's (1971) five-volume *The Philippine Insurrection against the United States* (PIR). It asks the preliminary methodological question: what precautions must a historian take when using the research software TM to understand the implicit criteria governing the creation of an archive?

This paper first introduces the PIR and then briefly grounds this note in two historical methodologies: the interrogation of archive formation and quantitative textual analysis. It then describes the principles behind TM and acknowledges some practical limitations to the present experiment. The body of this paper follows a discovery structure. It narrates this study's trials, errors, and adjustments to give readers a sense of methodological cautions when applying TM. Its experiments suggest that TM's application remains reliant on material that cannot practicably be included in the corpus for processing—the researchers' own accumulated knowledge. This research note's conclusion proposes mitigating this problem by including the least nebulous of these external materials into the processed dataset, particularly materials penned by the archivist in question, a suggestion that has yet to be properly tested.

Taylor's PIR

John R.M. Taylor curated two related corpuses relevant to the history of the Philippine revolution. The first is a manuscript collection composed of over 200,000 documents on the indigenous resistance movement collected by the United States military during the Filipino-American War.[1] Today, the National Library of the Philippines houses both physical and microfilm copies of this collection. The second corpus is Taylor's published selection of 1,534 documents taken from the larger collection. This smaller subset was produced from 1902 to 1906 in the context of the US military's need to justify their actions during the Filipino-American War. This five-volume set includes prose chapters by the author in volumes one and two. Nevertheless, the published work is mainly a compilation of transcribed and translated primary sources. Digitized (Optical character recognition [OCR]

processed) copies of these volumes have been produced by the National Historical Commission of the Philippines. The present research note examines this second corpus.

Articles by the historians John Farrell (1954) and John Gates (1985) best discuss the history of Taylor's published selection. To summarize, despite being initially supported by the government, Taylor's manuscript was eventually suppressed. The analyses Taylor produced were found to be politically inconvenient. The negative input of James Leroy, a rival scholar, further sealed the collection's fate. Taylor's volumes were only published in the 1970s funded by the Lopez Foundation and edited by Renato Constantino. Since then, Taylor's volumes have proven invaluable to scholars as a source of documents on the Katipunan, on the Republic, and on the wars that both groups were involved in (Linn, 2000; Rafael, 1995; Clymer, 1983).

It is not clear what criteria Taylor used when selecting his sources. He says that "the original documents upon which I have drawn form a collection of over 200,000 papers . . . Most of them are of no interest. To find what was [of interest], has, however, required an examination of all of them" (Taylor, 1971). How exactly Taylor determined what was of "interest" to US-Philippine relations is left to speculation. John Larkin (1976), for instance, has remarked that "while Taylor selected judiciously, he tended to favor those documents of national and international importance. Much material pertaining to provincial- and village-level affairs lies buried in the confusion of the unused portions." Apart from rudimentary indices made available by the National Library,[2] I have found no work that introduces the logic with which Taylor segregated his documents. Taylor cannot be faulted for his lack of clarity. We are, after all, dealing with his manuscript, not his final product. Nevertheless, an uninformed reliance on Taylor's unspoken criteria is risky.

Interrogating the archivist

The present study's aim is akin to that of Gloria Cano in her groundbreaking 2008 essay *Blair and Robertson's The Philippine Islands, 1493–1898: Scholarship or Imperialist Propaganda.* Cano first highlights the widespread use of the Blair and Robertson collection's (BR) 55 volumes in Philippine historiography. She then

uses the correspondence surrounding the production of these volumes as well as her knowledge of nineteenth century Philippine colonial history to highlight selections and omissions made by the collection's editors. Cano argues that the collection's composition is suggestive of these same editors' political motives. She concludes that the collection is imperialist propaganda.

In the collections of sources such as the BR and PIR, the editor is inevitably an archivist curating sources according to, often, undetailed priorities (Mojares, 2013). In Cano's article, her interrogation of the BR's archivists rests on crucial data external to the archive. Can a similar analysis be done for the PIR despite the absence of an equivalent database of damning correspondence between archivists? Careful textual analysis of the archivist's published collection, when compared and contrasted with his wider manuscript collection, may give us sufficient access to the archivist's priorities. But how does one efficiently undertake the analysis of large collections of 1,534 and over 200,000 documents respectively? A digitally supported quantitative approach may be the answer.

Counting words

A quantitative approach to textual analysis has some grounding in recent Philippine historiography. Benedict Anderson's 2003 and 2006 essays on the forms of consciousness found in Jose Rizal's novels are an early example. Combined in Anderson's (2008) volume *Why Counting Counts*, these essays systematically quantify the vocabulary of Rizal's novels to highlight his emphases and silences. A comparison between those emphases and silences, on the one hand, and the wider context of ideas to which Rizal was exposed on the other, is used to highlight Rizal's intentions. Anderson, for instance, argues that Rizal virtually omits references to Chinese mestizos as a social group in these novels, despite the fact that he "was perfectly aware of its existence and importance" (Anderson, 2008). This omission "suggests an intention to blur any distinction between the two main types of mestizo [*mestizo Chino* and *mestizo Español*]" (Anderson, 2008). To Rizal "they are all 'mixed,' all Catholic, all Spanish-speaking, all privileged: above all, *not foreign*" (Anderson, 2008).

Anderson's quantitative effort was done manually (2008). Subsequent advances in technology have allowed later scholars to undertake the effort digitally. Ramon Guillermo's 2014 article *Translation as an Argument: The Nontranslation of Loob in Ileto's Pasyon and Revolution*, for instance, applies this method to translation studies.[3] Asserting that the meaning of words rests on how they are used, Guillermo tallies the usage of the word loob across several colonial-era texts including the 1814 *Pasyong Pilapil*, a compilation of Andres Bonifacio's texts from 1894 to 1897, and several secular nationalist texts "deeply influenced by the Katipunan revolutionary idiom" (Guillermo, 2014). Ultimately, Guillermo systematically finds an overlap in how the Tagalog word *loob* is used in the *Pasyong Pilapil* and in Bonifacio's writings. However, a broader quantitative consideration of the vocabularies used by the above corpus suggests that the Bonifacio's text was lexically distant from the mystical *Pasyon* and lexically closer to the secular nationalist novels. In short, Guillermo argues that the historian Reynaldo Ileto's understanding of Bonifacio's concept *loob* solely in light of the *Pasyon* is more restrictive than helpful.

While both Anderson and Guillermo employ a quantitative approach to the close reading of a handful of texts, presently available software allows the researcher to easily scale up the size of analyzable datasets. TM software, which has been used in other fields to examine datasets as large as 12,500 and 13,300 documents, is one such option (Hall, Jurafsky, and Manning, 2008; Wang & McCallum, 2006).

Topic Modeling

The idea behind TM is that when people write, they tend to use certain sets of words alongside a subject. These words tend to occur together whenever people discuss that subject. To apply that idea in reverse: if one can find out which words usually occur together in a dataset, one can assume that these words refer to a common Topic. One will then be able to assume that the simultaneous presence of these words in a given document, when detected, signals the discussion of that Topic in the document (Brett, 2012). To reiterate, once a general per-Topic assignment of words is known, observable variables (the words of each document) can be used to arrive at a corpus' hidden Topic

structure (ex: the total Topics discussed in that corpus, the distribution of Topics per document in the same, and the distribution of documents per Topic in the same) (Blei, 2011).

The automation of this process is geared towards the analysis of the recurrent themes of a body of works too large (usually in the thousands) for an individual scholar to efficiently study on his/her own. It is not expected that TM will replace the role of researcher as the final arbiter of questions posed to and interpretations drawn from the data. TM is a tool. The present research note explores the ways that this tool's function will have to be determined by its wielder.

Scope and limitations

Ultimately, as with the articles of Cano and Anderson, knowledge of the archivist's intent will come from comparing and contrasting his/her archive with the wider variety of references that were available to him/her. While that is likewise the long-term goal of this research project, the present research note is a limited but necessary preliminary test of the utility of TM via the program MALLET for this task. In other words, the comparison of the themes of Taylor's selection (1,534 documents) against the themes of his total corpus (over 200,000 documents) is beyond the scope of the present project. As a preliminary test, this project processed only an initial 107 documents from Taylor's published collection to evaluate the feasibility of later applying TM to Taylor's much larger published and manuscript collections.

Limiting the sample to 107 documents kept this experiment manageable. Trial and error filled preliminary work. Naturally, (i) the cleaning and re-cleaning of a handful of texts to master a process before (ii) dealing with the issues of an entire corpus, seemed much more efficient than attempting to clean an entire corpus on the first try. The 107 documents were taken from volume 1 of Taylor's work. Chapters 3 to 5 of Taylor's prose analysis from this volume were also appended to the corpus during processing. These prose chapters are, at the moment, the closest glimpse Taylor gives of his intentions. At the same time, being included in Taylor's volumes, they are arguably inseparable from his corpus. Their inclusion in the dataset hopefully led to a more

nuanced automatic generation of topics.[4] In total, 110 documents were processed by this study although only 107 of them actually belonged to Taylor's corpus.

Method

The technical details of how TM uses the process called Latent Dirichlet Allocation (LDA) to arrive at the per-Topic assignment of words seen in Sample 1, and how LDA gets from there to Topic structures, are best tackled by the introductions by Scott Weingart (2012) and by Matthew Jockers (2011). These materials should be read with the video presentation *The Details* by David Mimno (2012). The present paper applies the software MALLET, developed by Mimno, to process its data. MALLET, on Windows, is a TM program operated via command lines inputted on the program PowerShell. How MALLET is used, what input and parameters it employs, and what output one can expect from it is explained below.[5]

Over the next few pages, I take the reader though a detailed but nontechnical description of the two rounds of TM to which I subjected my dataset, along with the changing parameters of data organization and cleaning that facilitated this process. Most importantly, for each round of TM, I evaluated the results of manual and automatic codebook generation and code assignment.

Data organization and cleaning

TM generally has two output files. The first is a list of all the automatically generated strings of co-occurring words (see Sample 1 below) where each string represents a Topic. The second is a chart of Topic assignments per document where each row represents one document used, and every column alternates between Topics and their associated rank of relevance to the document. Above average scores of relevance indicate that the software is confident in assigning a given topic to a given document. The values on these columns are arranged in descending order from left to right (see Sample 2 below).

SAMPLE 1: List of automatically generated Topics

Topic #	Strings of co-occuring words
0	artillery companies battalion regiment infantry soldiers officers heavy military cazadores replace jan total regiments mountain disciplinary present number marine
1	desire public instruction kinds acts industry nation industries institutions rights laws north protection soil countries legitimate law armistice counting
2	clergy secular regular priest chinese priests parishes parish land change estates hands jesuits ecclesiastical bishops mestizos increase changed america
3	natives islands spanish native years religious laws priests church law administration work great system orders part force hundred peace
4	general de aguinaldo primo biac arms na rivera paterno spanish bat government peace surrender war agreement december spain signed
5	general people manila country time civil orders authorities archipelago called philippine islands authority filipino power made philippines cavite great
6	rizal league propaganda association committee finally pilar dapitan solidaridad madrid morayta lodges founded barcelona formed orient pamphlets santiago head
7	article government secretary president council constitution office assembly republic army representatives treasury secretaries foreign central charge revolution interior vote

SAMPLE 2: Chart of relevant Topics per document

Document file name	Topic #	Rank in relevance	Topic #	Rank in relevance	Topic #	Rank in relevance	Topic #	Rank in relevance	Topic #	Rank in relevance
1897_01_12_E_64.txt	4	32%	16	16%	33	12%	11	9%	18	7%
1897_08_00_E_40.txt	39	41%	4	18%	54	13%	5	13%	22	5%
1897_09_03_E_43.txt	14	32%	23	24%	10	8%	22	8%	7	6%
1897_09_20_E_45.txt	4	34%	14	20%	10	10%	33	6%	11	5%
1897_11_15_E_53.txt	4	33%	54	30%	28	28%	41	2%	11	1%
1897_11_18_E_55.txt	52	30%	4	22%	5	11%	33	7%	10	6%
1897_12_14_E_59.txt	4	36%	54	30%	41	5%	33	5%	10	5%
1898_01_24_E_72.txt	9	28%	4	18%	11	12%	12	10%	14	5%

1898_09_15_E_69.txt	54	17%	4	16%	53	10%	10	7%	22	6%
1899_01_12_E_84.txt	45	26%	42	17%	22	12%	58	10%	34	7%

Each of the 110 documents was saved as a notepad file. They were labeled chronologically as follows: "YYYY_MM_DD_[Chapter # or Exhibit # example: 'E_04']." All missing dates were replaced by "0." The goal here is, in future experiments, to take advantage of the metadata of time.6 Labeling files by date makes it easy to chronologically arrange the files' data on an Excel spreadsheet and to create time-based charts from the results of data processing.

Taylor sometimes combined several documents under one exhibit number but distinct item numbers. Such was the case with document 11, which Taylor actually divided into 11a to 11i. Since Taylor seemed to have wanted to treat such documents as distinct, the present project treated all such documents as individual documents.

In other cases, Taylor treated multiple documents as inseparable and no distinct item number was assigned to each additional document. For instance, Taylor kept the two documents of Exhibit 5 under the label "5." Suspecting that Taylor treated these documents as possessing a single idea, the present project retained all such documents as single documents. When two documents meant to be combined were dated differently, their combined file was saved under the filename appropriate to the later document's date.

Initial cleaning undertaken for this project took three steps:

1) The removal of Constantino's brackets over words that Taylor spelled incorrectly (ex: "Katipunan" edited from "Katip[ún]an");

2) The correction of words that OCR processing misread, such as accented words (ex: "Katipunan" edited from "Katiptian"); and

3) The correction of words broken into two either i) by faulty OCR processing or ii) because a long word reached the end of a line, broke at the margin, and then continued in the next line.

TM Round 1

An initial round of TM yielded the following strings of co-occurring words (see Sample 3). Included in that chart (left - most column) are my preliminarily best-guess labels for these strings.

Sample 3: Resulting list Topics from initial Topic Modeling

Working Title	Topic #	Strings of co-occuring words
US funds Aguinaldo's return	3	aguinaldo united states hongkong philippines consul spanish general april insurgents american payment agreement junta singapore dewey mr money pratt
Spanish docs	11	de la el en los sr por se las filipinas su con al president del es para presidente philippines
Payments made by Spain	16	manila spain made governor de part philippines held leaders found considered end fz left house gave possession provinces paid
Bonifacio's trial	17	bonifacio andres government rifles soldiers troops brothers men day house aguinaldo court twenty army signed judge witness thousand town
Katipunan military administration	19	art military chief number provisions proper order council officers commander enemy report president company mentioned officer decree information article

Sample 4: Resulting chart of relevant Topics per document from initial Topic Modeling

Document file name	Topic #	Rank in relevance	Topic #	Rank in relevance	Topic #	Rank in relevance	Topic #	Rank in relevance
1898_05_04_E_91.txt	11	72%	3	5%	5	5%	8	4%
1897_05_08_E_30.txt	17	71%	8	10%	19	5%	16	4%

A comparison between these hypotheses and the documents they described displayed mixed results. Document 91's assigned Topics were fairly accurate. This document was a minutes of a meeting in which the Hong Kong junta of revolutionary exiles deliberated on whether or not their leader, Emilio Aguinaldo, should agree to meet with American representatives to facilitate the exiles' return to the Philippines. The junta weighed the risk that the US might use Aguinaldo's prestige to later colonize the Philippines. Taylor published an English and a Spanish version of this document. Appropriately, my first round of TM coded the document as "US funds Aguinaldo's return" (Topic 3) and "Spanish documents" (Topic 11).

Meanwhile, document 30's assigned Topics were off the mark. Document 30 discussed the Katipunan's military organization and recruitment as well as contributions. It was coded under my labels: "Bonifacio's trial" (Topic 17), "Katipunan military administration" (Topic 19), and "Payments made by Spain" (Topic 16). Only the second of these Topics was appropriate.

In general, it seems that even if each TM string lists words that commonly co-occur together in most documents across all documents, only a few of these words need ever appear together in any one document for TM to detect co-occurrence. In other words, not all of a co-occurring string of words such as "*orange, apple, pear*, and *mango*" need to show up in any one document for TM to detect that these terms generally co-occur in the dataset.

Moreover, my Topics overlapped too greatly with one another. Several distinct ideas were combined in single strings of words. This combination was understandable since several ideas (and so also the words associated with these ideas) may indeed have co-occurred in the same text. For example, the words "mango" and "pear" (more usefully categorized as "fruits") might often co-occur with the words "knife" and "fork" (more usefully categorized as "eating utensils"). Ideally the researcher will hit upon the category "snack" under which both sets of words fall under. However, if the combination of many smaller ideas is too broad, it runs the risk of garbling the resultant TM strings, making it difficult for the researcher identify any of the distinct Topics present.

Revision of parameters

To remedy the latter problem, I decided to add new parameters to my command line, changing its default settings. Instead of asking MALLET to produce the default # of Topics (20 Topics), I asked it to produce 60 Topics. Also, instead of asking it to run the default # of iterations, I asked it to run 2000 iterations.[7] I adopted both strategies to facilitate the detection of (a) Topics that appeared in only a few documents but were greatly relevant each of document, rather than (b) Topics found in many documents but that were only slightly relevant to each document (Mimno, 2012).[8] In short, the process would reduce overlaps between generated Topics, making results more distinct/specific and so more accurate.

The resulting command line, with the revised parameters in bold font, is detailed below:

> PS C:\mallet> bin\mallet train-topics —input digmeth200115.mallet —num-topics 60 —num-iterations 2000 —optimize-interval 20 —output-state DGR2T60I2000-state.gz —output-topic-keys digmeth20015R2T60I2000_keys.txt —output-doc-topics digmeth200115R2T60I2000_composition.txt

In addition, I decided to add Spanish stopwords to the default English stopwords (Ranks NL company, n.d.). These stopwords reduced the appearance of the words “de” and “la,” which frequently appeared in surnames, and so in resultant TM strings, but were irrelevant to my research goal.

TM Round 2: Codebook generation, manual vs. automatic

To evaluate the Topics generated during this second round of coding, I took the following steps:

1) I manually created a codebook for the qualitative data of a selected number of documents (chapters 3 to 5, and exhibits 1-11, 21-22);

2) I examined the strings of words (generated Topics) automatically associated with each of these documents at a relevance of at least 10%; and

3) I attempted to see if my manually generated codebook matched the automatically generated list of Topics.

The results of these initial steps were disconcerting. I manually detected and assigned forty-one (see Sample 5) different Topics to these documents, but only nine (see Sample 6) fit TM's automatically generated Topics.

Sample 5 : Manually generated codebook

CODEBOOK - PIR - TAYLOR. Taken from chapters 3 to 5, and exhibits 1-11, 21-22	
C001-200	TEXT ALONE - SUBNET
C001-020	ORGANIZATIONS - SUBNET
C001	Masonic lodges
C002	La Liga
C003	Secret organizations gen.
C004	Rivalry
C005	Colonial government
C006	Religious orders
C007	American government
C021-040	ORGANIZING EFFORT - SUBNET
C021	Organizational structure (officers etc.)
C022	Recruitment
C023	Financing and management of funds (taxation etc.)
C024	Leadership (elections?)
C025	Statement of ideals, goals and responsibilities (Oath taking, call to arms)
C041-060	CHARACTER OF PEOPLE INVOLVED - SUBNET
C041	Irrational
C042	Brutal, savage
C043	Primitive
C044	Deceitful (ex: employing vagueness in order to make later claims)
C045	Anti-Friar
C046	Plebian
C061-080	PERPCEPTIONS ABOUT THE RELIGIOUS ORDERS - SUBNET
C061	Manipulating civil government
C062	Positive things the priests have done
C063	Good intentions
C064	They generally go unpunished for crimes
C065	They damage the reputation of Spain
C066	They oppose secularization / Secular priests are against them
C067	The claims against them are made up / unjust
C068	They extort money (excessive dues and rents)

C081-100	MILITARY OPERATIONS - SUBNET
C081	Reinforcements from Spain
C082	Infrastructure and transportation (destruction, creation)
C083	Filipinos loyal to Spain
C084	Entrenchment and fortifications
C085	Logistics of supplying defense
C086	Procurement of arms
C087	Deportations and trails of conspirators
C101-120	DIPLOMACY AND NEGOTIATIONS - SUBNET
C101	Negotiations with the USA
C102	Planned return to the Philippines
C103	Surrender at Biac na Bato (objections, actions etc.)
C104	Biac na Bato constitution
C105	Question of promised reforms by Spain
C121-140	USE OF THE SURRENDER MONEY - SUBNET
C121	Future investment in the war
C122	Discontent of leaders who did not get a share
C123	Process of remittance of money

Sample 6: Topics from Sample 5's manual assignment that observably coincide with MALLET's automatically generated Topics.

C201-220	FRIARS - SUBNET
C201	Action the revolution is taking against them
C202	Actions they are taking against the revolution
C203	Negative acts they have done against the people (pre-revolution)
C221-240	HONGKONG JUNTA - SUBNET
C221	Junta's managament of surrender money
C222	Negotiations with the USA
C223	Pact of Biac na Bato
C241-260	MOBILIZATION
C241	Statement of ideals and goals and responsibilities (Oath taking, call to arms)
C242	Organization of local government
C290-C300	OTHER
C291	Written in Spanish

TM's generated strings combined some of the Topics that I considered to be about two distinct themes into single Topics. For example, TM combined the Topics (a) "Negotiations with the

USA," and (b) "Junta's management of surrender funds," which I perceived as distinct, into the single string (c) "*aguinaldo hongkong united states leaders insurgents philippines agreement junta insurgent payment april consul funds bank january deposit commodore arrived*."

This issue may stem from a more generally applicable rule. The researcher seems to need three sets of criteria to derive Topics from TM's results: 1) the researcher's own implicit criteria, which are interfaced with a body of stock knowledge that not even he or she is fully conscious of; 2) the researcher's conscious criteria, which are interfaced with his/her bibliography's combined knowledge of historical context; and, 3) the criteria of TM, that is, co-occurrence. These criteria seriously affect how well results derived independently by manual and automatic coding can coincide. MALLET links words together because of co-occurrence. If manual coding operates in a manner other than co-occurrence, then TM cannot be expected to replicate the results of manual coding.

The software may split a matter that the historian considers a unitary topic into two separate topics. This split might happen if MALLET, when sifting through words, makes a particularly fine distinction between strings that is irrelevant to the researcher's study. In this way MALLET treated both strings (a) "*nan katip society brothers jesus blood swear true members brother lord life punishment drop hour oath accord punished membership*" and (b) "c*ountry honor noble good worthy free beloved love association principles citizens respected attain knowledge patriotic language false hearts exception*" as distinct. But to the researcher both strings may express nothing other than the unitary Topic "C025 Statement of ideals and goals and responsibilities". During the evaluation of TM's results, knowledge external to the corpus and criteria other than co-occurrence seem crucial to judging when a topic is relevant.

A generated string of words may look like nonsense to the historian, simply because its co-occurrence follows an internal logic unfamiliar or irrelevant to the historian. For example: the string "*senor excellency spanish paterno estella reforms paternal pardon marques generous noble honor satisfaction full revolution heard alejandro*

religious love", can fall under the general header "praises," which may be irrelevant to an economic historian.

Lastly, what degree of co-occurrence is actually necessary for the historian? Low-level co-occurrence, low enough to go undetected during TM's automatic coding, may nevertheless be relevant to the historian due to the historian's broad knowledge of context external to the corpus.[9]

TM Round 2: Code assignment, manual vs. automatic, independent code assignment

In this next test, I focused on only Topics that were commonly detected by both manual and automatic processes in the previous subsection (so Sample 6 only). My intention was to see if the strings automatically assigned to particular documents proved to be a useful way of gauging the topics that those documents contained. In other words, if an automatically generated string included the words: "orange, peach, pear, and apple," a string I would entitle "fruits," then, ideally, both manual and automatic code assignment will detect the very same documents as having the topic "fruits." To avoid biasing my reading, I undertook manual code assignment before the automatic generation of Topic strings. Sample 7 below shows that manual and automatic results coincided, but only to an imperfect extent.

Sample 7: First round comparison of manual and automatic assignment Topics

		Manual Assignment			Automatic Assignment	
Document/ Exhibit	1	C242			C242	
	2	C203			C202	
	3	C203	C203		C201	
	4				C201	C241
	5					
	6	C201			C201	
	7	C201				
	8	C203			C203	C201
	9					
	10	C242				
	11a	C241			C241	C201
	11b	C241			C241	C242
	11c	C241			C241	
	11d	C241			C241	

	11e				C223	
	11f					
	11g				C241	
	11h				C241	
	11i	C242			C241	C242
	21	C241	C242			
	22	C203				
Chapter	3	C201				
	4	C221	C223		C223	C222
	5	C221	C222	C223	C222	C223

TM Round 2: Manual code assignment, documents preselected for co-occurring words

I suspected that the above incidences of non-coincidence were caused by an unintended prioritization of (a) the Topics generated by the human-researcher's set of criteria (as discussed in the first section of TM Round 2 above), over (b) the TM criteria of co-occurrence. So for the final test, I decided to mitigate this potential source of bias by reversing my code assignment process. In other words, I sought to prioritize TM's criteria of co-occurrence over the human-researcher's criteria. Rather than begin with manual coding, I began with 10 documents that MALLET detected as having an above average applicability of Topics from the Sample 6 codebook.10 These 10 documents were then manually coded.

As suspected, the results (see Sample 8) demonstrated a much higher coincidence between manual and automatic Topic assignments per document than in Sample 7. However, the prior knowledge that these 10 documents were guaranteed to fall under the same automatically assigned topic may have ended up forming a part of the stock knowledge with which I undertook manual code assignment. In other words, I may have biased my manual coding into the topics generated by TM.

Sample 8: Second round, comparison of manual and automatic assignment Topics

		Manual Assignment		Automatic Assignment	
Exhibit	40	C223		C223	
	43	C242		C242	
	45	C223		C223	C291
	53	C223	C291	C223	
	55	C223	C221	C223	
	59	C223		C223	C221
	64	C223		C223	
	69	C221	C223	C223	
	72			C223	
	84	C222		C222	

A dilemma then confronts me. Should one undertake close reading first and use TM's test of co-occurrence only to confirm the results of manual coding, or run the corpus by TM, examine the resultant strings, and then massage them afterwards, combining and dividing them into Topics via close reading? The first option has its value, but if the goal of using TM is to increase efficiency in tackling a large corpus, then the first option defeats this purpose by requiring time consuming close reading of every document. However, the second option is not necessarily better. Knowing that the computer is assigning codes based on nothing but co-occurrence, it may be risky for researchers to use the computer's Topic assignments as the starting point of their analyses. The researcher risks being distracted by software's limitations.

Concluding observations

This research note set out to conduct and document a preliminary test of TM towards a greater goal of interrogating the archivist. It examined a selection of 107 documents taken from a corpus. The data was cleaned twice: initially to proof read the OCR data and then later to deal with more distinct and so easily recognizable topics and to include Spanish stopwords.

Three comparisons were then run between manual and automatic coding efforts. First, I generated codebooks both

manually and automatically. The results were very different. Second, I coded the corpus based only on codes found both on manually and automatically generated codebooks, beginning first with manual coding before automatic coding. The results coincided but imperfectly. Then I reversed the process by manually coding only documents that TM first found to have the same topic. While these results coincided, the risks of beginning an analysis based on co-occurrence, rather than on manual coding, remained ominous.

These preliminary tests helped clarify the gap between manual and automatic coding. Because manual coding employs data external to the corpus while automatic coding focuses on co-occurrence solely within the corpus, it seems that the latter cannot be expected to replicate the former. In terms then of this project's specific goal of interrogating the archivist, if one resolves to exploit TM for its efficiency, whatever its limitations, it might be helpful to remember that co-occurrence is useful only when the historian is able to use material external to the corpus to distinguish between relevant and irrelevant co-occurrences. It would also be helpful to explore to what degree material completely external to the corpus but highly relevant to the archivist's point of view (e.g. separately published political treatises written by the archivist) can be incorporated into the corpus. Future tests will need to gauge whether that inclusion would, on balance, result in a study that is over-compromised by a bias in the TM due to the external material selected by researcher for inclusion into the corpus, or a study that best surfaces co-occurring sets of words relevant to researcher's analyses and closest to the archivist's actual priorities.

Acknowledgements

I thank Peter Xenos, Ariel Lopez, and the two anonymous referees whose encouragement, comments and suggestions guided the revision of this essay. I also thank the National Historical Commission of the Philippines for allowing me access to their library data bank.

Nicholas Sy is an Assistant Professor at the University of the Philippines, Diliman, Department of History. In 2017 he received his Master's degree in History at the Ateneo de Manila University. For his MA thesis, Nicholas presented a quantitative challenge to the common notion that colonial era political elites primarily married within their own social circle. His published essays include "Horacio de la Costa, Foreign Missionaries, and the Quest for Filipinization: The Church in the Age of Decolonization," which was co-authored with Filomeno V. Aguilar, Jr. and published in Philippine Studies: Historical and Ethnographic Viewpoints. His research interests are in church history, demographic history, and quantitative approaches to history.

Notes

[1] This collection has been subject to neglect, both by its early American documenters and by The National Library of the Philippines. The latter, has, for example, allowed one of its copy's most important reels, the rudimentary table of contents, to succumb to vinegar syndrome.

[2] These materials were, until a recent system crash, made available by The National Library of the Philippines on their website's digital collection section, under the title: "Philippine insurgent records," and the specific header: "guides." Taylor and the microfilmers that inherited his collection each attempted preliminary and partial indices (*Philippine Insurgent Records*, p. 3–4, 7–8, 11). Neither attempt made it to print. The indices produced by Taylor himself for his book were the more important (*Philippine Insurgent Records*, p. 3–4, 8, 11). However, many of these materials are now missing. The microfilmers of his collection later complained repeatedly about the "archival chaos" (ibid., p. 20) that Taylor had left his collection in.

[3] See also Guillermo's 2009 book *Translation and Revolution: A Study of Jose Rizal's Guillermo Tell*.

[4] With this same logic I used only the documents of volume one (rather than take samples from across the volumes) to allow for a closer interfacing between the themes of volume one's documents and the themes of volume one's own prose chapters.

[5] A good introduction on how to use MALLET towards TM can be found in a blog post by Shawn Graham, Scott Weingart,

and Ian Milligan (2012) called *Getting started with topic modeling and MALLET*.

[6] Other metadata to consider would be the ethnicity of the writer, the length of the text etc. The above and other metadata can be prepared for tracking-over-time by making similar adjustments to the filename based on each category and its subcategories (so for example: all documents written by Spaniards can have the additional tag "1" in their filenames and all documents written by Americans can be tagged "2").

[7] Iterations are the repetitions of MALLET's process of contextualizing every word within its co-occurring words.

[8] See also Annie Swafford's (2014) class blog on digital history *Digital Tools: Sherlock Holmes' London*.

[9] On a related note, one wonders how synonyms are accounted for. If every document mentioning fruits uses a different synonym of spoilage (ex: spoiled, rotten, worm-ridden) then it sounds likely that the topic of fruits may not co-occur with any of the words relevant to the topic of spoilage, even if references to spoilage are prevalent across documents.

[10] See sample 2 for a reminder of what the ranks of relevance mentioned here are.

[11] Based on internal evidence, this document was written after 1957, probably around the time the microfilmed copies of the P.I.R. were produced and shipped to the Philippines.

References

Anderson, B. (2008). *Why counting counts: A study of forms of consciousness and problems of language in Noli me Tangere and El Filibusterismo.* Quezon City: Ateneo de Manila University Press.

Blei, D. (2011). Introduction to probabilistic topic models. Retrieved December 2014, from https://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

Brett, M. R. (2012). Topic modeling: A basic introduction. *Journal of Digital Humanities*, *2*(1). Retrieved January 2015 from http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Cano, G. (2008). Blair and Robertson's *The Philippine Islands, 1493–1898*: Scholarship or imperialist propaganda? *Philippine Studies,* 56(1), 3–46.

Clymer, K. J. (1983). Review of *Benevolent assimilation²: The American conquest of the Philippines, 1899–1903. American History*, 11(4), 547–552.

Farrell, J. T. (1954). An abandoned approach to Philippine history: John RM Taylor and the Philippine insurrection records. *Catholic historical review*, 39(4), 385–407.

Gates, J. M. (1985). The official historian and the well-placed critic: James A. LeRoy's assessment of John RM Taylor's *The Philippine insurrection against the United States. The Public Historian*, 7(3), 57–67.

Graham, S., Weingart S., and Milligan, I. (2012, September 2). Getting Started with Topic Modeling and MALLET. [Web log post]. The programming historian. Retrieved December 2014 from http://programminghistorian.org/lessons/topic-modeling-and-mallet

Guillermo, R. (2014). Translation as an argument: The Non-translation of Loob in Ileto's *Pasyon and Revolution*. *Philippine Studies: Historical and Ethnographic Viewpoints,* 62(1): 3–28.

Guillermo, R. (2009). *Translation and revolution: A study of Jose Rizal's Guillermo Tell*. Quezon City: Ateneo de Manila University Press.

Hall, D., Jurafsky, D., and Manning, C. (2008). Studying the History of Ideas Using Topic Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Association for Computational Linguistics. Retrieved August 2018 from https://dl.acm.org/citation.cfm?id=1613763

Jockers, M. (2011, September 11). The LDA buffet is now open; or, Latent Dirichlet Allocation for English majors. [Web log post]. Matthew L. Jockers. Retrieved December 2014 from http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/

Larkin, J. A. (1976). Review of *The Philippine insurrection against the United States: A compilation of documents with notes and introduction*. *The American Historical Review*, 81(4), 945–946.

Linn, B. M. (2000). *The Philippine war, 1899–1902*. Lawrence: University Press of Kansas.

Mimno, D. (2012). The details: How we train big topic models on lots of text. From MITH Topic Modeling Workshop, Maryland, 3 Nov. [Instructional Video]. Retrieved January 2015 from http://vimeo.com/53080123

Mojares, R. (2013). *Isabelo's archive*. Mandaluyong City: Anvil Publishing.

National Library of the Philippines. [n.d.] Philippine insurgent records. [Digital Collection]. Retrieved January 2015. http://nlpdl.nlp.gov.ph:9000/rpc/cat/finders/PI01/guides.htm

Philippine Insurgent Records, 1896–1901 with associated records of the United States war department, 1900–1906. [Digital Collection]. Retrieved January 2015 from http://nlpdl.nlp.gov.ph:9000/rpc/cat/finders/NL02/NLPPIGD20101107114/date.htm[11]

Rafael, V. L. (1995). Colonial domesticity: White women and United States rule in the Philippines. *American Literature*, 67(4), 639–666.

Ranks NL company. [n.d.]. Spanish Stopwords. [Web log post]. Ranks NL webmaster tools. Retrieved January 2015 from http://www.ranks.nl/stopwords/spanish

Swafford, A. (2014). Digital tools: Sherlock Holmes's London. [Weblog]. Retrieved January 2015 from https://sherlockholmeslondondh.wordpress.com/about/

Taylor, J. R. M. (1971). *The Philippine insurrection against the United States: A compilation of documents with notes and introduction* (Vols. 1–5). R. Constantino (Ed.). Pasay: Eugenio Lopez Foundation.

Wang, X. and McCallum, A (2005). Topics over time: A non-Markov continuous-time model of topical trends, presented at Conference on Knowledge Discovery and Data Mining, Philadelphia, 2005.

Weingart, S. (2012.) *Topic modeling for humanists: A guided tour*. Retrieved from http://www.scottbot.net/HIAL/?p=19113. Accessed December 2014.