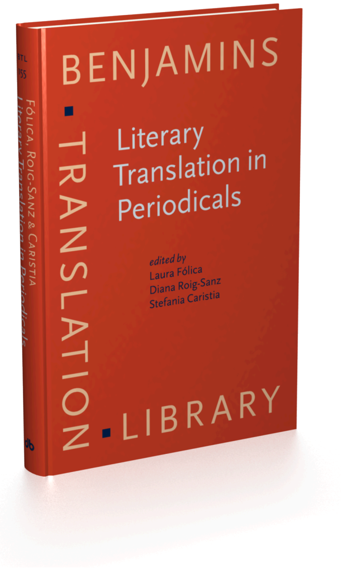
Benjamins Translation Library
155

CHAPTER 10

Challenges and strategies for beginners to solve research questions with DH methodologies on a corpus of multilingual Philippine periodicals

Rocío Ortuño Casanova | University of Antwerp

https://doi.org/10.1075/btl.155.10ort


BENJAMINS
TRANSLATION
LIBRARY
Literary Translation in Periodicals
edited by
Laura Fólica
Diana Roig-Sanz
Stefania Caristia
BTL
155
Fólica, Roig-Sanz & Caristia
Literary Translation in Periodicals

Pages 247–272 of
Literary Translation in Periodicals: Methodological challenges for a transnational approach
Edited by Laura Fólica, Diana Roig-Sanz and Stefania Caristia
[**Benjamins Translation Library**, 155] 2020. vii, 401 pp.

John Benjamins Publishing Company

CHAPTER 10

Challenges and strategies for beginners to solve research questions with DH methodologies on a corpus of multilingual Philippine periodicals

Rocío Ortuño Casanova
University of Antwerp

A usually mentioned problem in Digital Humanities (DH) is the difficult fit between Humanities research questions and DH methodologies. This chapter is therefore configured as a *meta-chapter* that explains the problems and strategies when exploring the multilingual repository of Philippine periodicals constructed within the project "Strenghthening Digital Research at the UP System" in order to research the evolution of the image of China in these periodicals. The two main challenges found for analysing the periodicals to find an answer have been (1) Problematic OCRs, (2) Research across multi-lingual publications. The chapter lists literature and research projects that have approached similar questions and challenges in comparable corpora. Some suggestions of tools to address them will also be provided.

Keywords: Philippine rare periodicals, multilingual text analysis, representation, low-resource languages, OCR, online repository, challenges in digital humanities

Talking about Digital Humanities in the Philippines involves entering a number of global debates either voluntarily or involuntarily. One of them is the whiteness of the discipline and the inclusion of southern countries (Crompton, Lane and Siemens 2016: 22–24). One of the greatest barriers in this sense is the fact that digital humanities are expensive. We work with digitised texts that need first to be digitised. In order to do that, salaries and machines are needed, but also time and training, often abroad. To analyse the results, more training, programs and computers are needed, and thus more money. Access to journals – by paying a fee – will also be needed to keep abreast of developments; results will have to be

https://doi.org/10.1075/btl.155.10ort

published and servers bought. If that sort of money already poses a problem in northern countries, where competition for obtaining project funding is fierce and frequently involves resorting to free labour from students, people on placements and volunteers for a number of tasks, in southern countries it presents one of the biggest obstacles and is probably one of the least mentioned issues in sessions on "Digital Humanities in the South" at the various Digital Humanities meetings held across the world.

The second debate is multilingualism. On the global stage, this debate has been questioning the hegemony of English in the Humanities for some years and promoting the search for ways to include, raise awareness and work with other languages. In the local sphere where we work the issue is a little different and makes us think about what Digital Humanities is trying to achieve. In this project, we will attempt different approaches to a multilingual set of texts and problem-solving strategies, including translation into a lingua franca. Although this is not an ideal solution, due to the current state of affairs regarding DH tools, it might be useful. Therefore, the problems approached will not be strictly translation-based, however, the translation will be a problem-solving strategy for this corpus.

The Philippines is not easy to locate conceptually speaking, due to its nature as a border country lying between the China Sea and the Pacific Ocean. For years it was considered part of Oceania; it was successively a Spanish colony, part of the Viceroyship of New Spain together with Mexico, the military district of Guatemala and of Cuba, a US colony and finally a Japanese colony. It is a country with no territorial continuity, constituted by between 7107 and 7600 islands (*GMA News Online* 2016). These geo-historical circumstances have posed a series of problems in both the study and the epistemological understanding of the country and its visibility abroad. In the first place, the diversity of geographical affiliations it has endured (Europe, America, Asia, Oceania) means that it has been pushed to the periphery of studies on these regions: Hispanist studies do not include it, and North American or Asian studies seldom approach it (Ortuño Casanova 2017: 59–60). Moreover, this situation has prompted constant questions about identity and history, the answers to which Digital Humanities could take great leaps forward in finding.

Therefore, if this is about understanding the cultural heritage and the past of a young country – independent since 1946 – based on its texts and other cultural products, there is no doubt that we should bear in mind that these cultural products will be expressed in one or other of the nation's 175 indigenous and 8 non-native languages, 41 of which are institutional languages. The official languages across the territory are Filipino and English, although much of the historical documentation is in Spanish ("Philippines" n. d.). This is the case of the historical periodicals that we are digitising and uploading to an online repository as part of the PhilPeriodicals project[1], developed by the ACDC group ('Antwerp Centre for Digital Humanities and Literary Criticism – ACDC – University of Antwerp' n.d.) and the University of the Philippines (Diliman) funded by VLIRUOS.[2]

This chapter is structured as a meta-study and a literary review in which different tools -easy to access and to master even if no specific training has been received on the matter- are suggested for their ability to cope with the difficulties surrounding multilingualism and the budgetary constraints of the repository being created with development cooperation funds. It also seeks to answer a specific research question by analysing periodicals in several languages simultaneously, as should be the case when, once the first part of building the repository is complete, the second part comes into action, namely digital humanities training for Filipino academics and implementing projects based over there. Those Filipino scholars are specialised in Humanities, without any prior knowledge of Digital Humanities, and they should be able after two summer schools of one week each, of setting up their own projects using digitised sources and DH methodologies. It needs to be taken into account that the methodologies selected assume that researchers and prospective researchers cannot program yet, and therefore, no programming is contemplated. The reason is that this is a first introduction to be tested after the first summer school of 1 week. After the second summer school and online training between both summer schools, students should be familiarised with Python.

This text also constitutes an assessment of what a researcher may need from a repository of historical periodicals in the Philippines and an initial look at the challenges to be tackled in the second phase of the project, when the platform starts being used for research purposes.

The PhilPeriodicals project

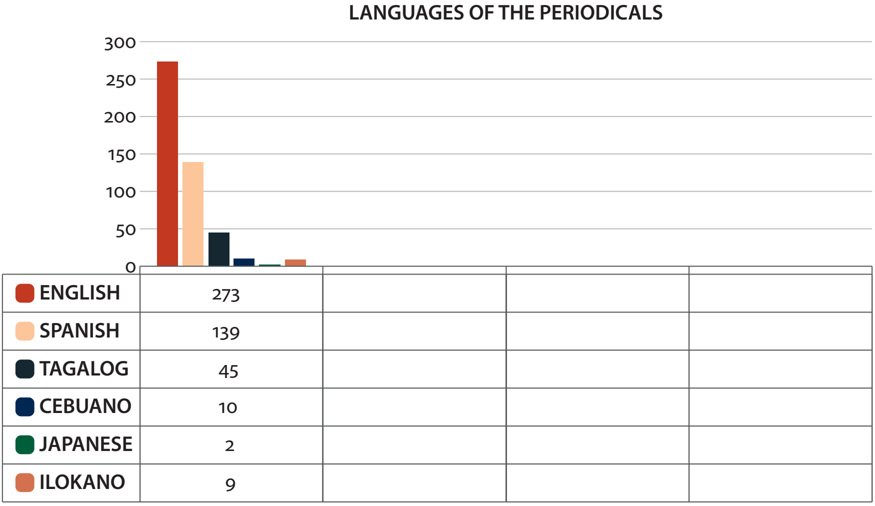
In April 2016 the Faculty of Arts and Letters at the University of the Philippines Diliman Campus was completely burnt down, along with a languages library containing dissertations and rare editions as well as a number of historic collections ('Fire Breaks out at UP Diliman Campus' 2016). In June 2018 another fire broke out in the main building of the historic archive of the Philippines in Manila ('Fire Hits National Archives Building' 2018). Although, according to press reports, the damage in the second fire was limited, it has not yet been possible to reopen the collection to

[1] Although it is still a work in progress, a first beta version of the repository can be consulted here: https://repository.mainlib.upd.edu.ph/

[2] https://hosting.uantwerpen.be/philperiodicals/

the public.[3] These are by no means isolated cases and they illustrate the precarious situation of documents that have survived adverse weather conditions and a world war that devastated the city of Manila. Moreover, in the case of the periodicals involved in this project, they are housed in a fire prone library (Lagrama 2012).

These materials are the ones we chose as "raw material" when thinking about a Digital Humanities cooperation project in the Philippines. What we call *Rare periodicals* is a set of 478 titles in six languages, some in more than one language simultaneously, and with an uneven number of issues for each title (Figure 1), published between 1862 and 1945 (Figure 2), that is, up to the end of World War II. The majority are incomplete collections, stored on paper and microfilm, and also of variable quality.



🟥 ENGLISH	273			
🟧 SPANISH	139			
⬛ TAGALOG	45			
🟦 CEBUANO	10			
🟩 JAPANESE	2			
🟫 ILOKANO	9			

Figure 1. Newspapers and magazines housed in the main library at the University of the Philippines by language

This library assessed as being "fire prone" holds other materials apart from periodical publications (books for example). This might prompt the question of why we chose such a mixed set of materials as a basis for the project. The reason is connected with the second phase of the project, which consists of organising a series of courses on Digital Humanities in the Philippines using the pyramid or cascade model. This means that one course is held in Manila and attended by representatives from other University of the Philippines outlying campuses (specifically Baguio, Iloilo and Cebu). Subsequently, those attendees and others from the first course go to each of those outlying campuses to teach courses similar to the first

[3] In January 2020 it had not been yet opened.

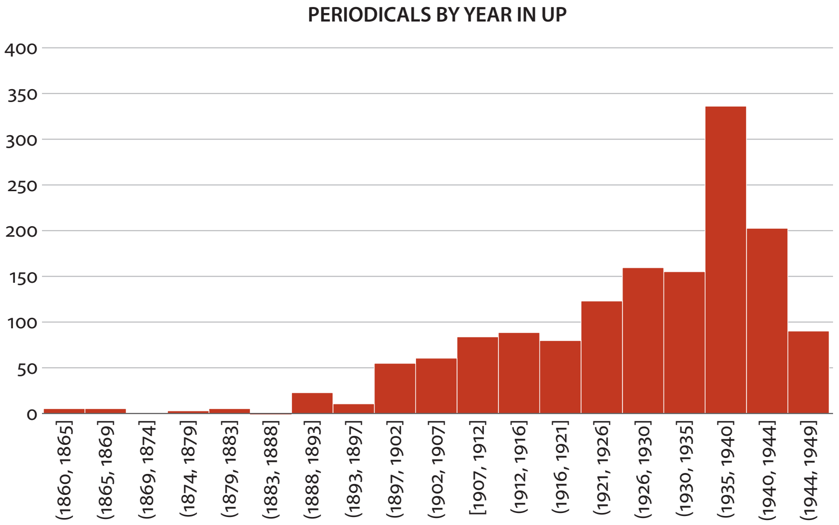


Figure 2. Number of periodicals housed in the University of the Philippines by year

one. It is hoped that this will enable several Digital Humanities research projects to be set up. As a consequence, one of the premises is to get as many researchers as possible involved. Periodicals can be useful as study material for historians, linguists (both in linguistics and literature and in cultural studies), sociologists, anthropologists and, if we include specialised journals (which we do), for agronomists, architects and a long list of others.

But in studying the periodicals field from the different points of view offered by journals and periodicals published simultaneously in different languages, we can also shed light on what the Philippines were like before 1945, on the ruling political forces and even on the Filipino literary canon, the rise and fall of languages, and international relations. Consequently, this process will highlight the position of the archipelago in the international complex of relations, and lead research to the way towards a "counter-hegemonic" globalisation (Li 2003: 11, 14–18).

What Pierre Bourdieu called the journalistic field is part of the field of cultural production and is seen as part of the field of power. Therefore, it interacts with the political field and with the field of social sciences. As a result, studying it will provide us with a broad view of how forces interact through cultural production in the country (we are thinking here of the writers who are also editors of one or more periodicals, who publish their stories in them, and how this expands knowledge of the literary field) (Benson and Neveu 2005: 5). In terms of literature, most Filipino literature from the first decades of the 20th century was published in printed periodicals. Thus, through the proposed research, it will be studied in the context and in comparison with other literatures. To sum up, by going back to the newspapers and magazines we will be able to reconstruct a cultural history of the Philippines with no linguistic intermediaries and colonisers, and in context and connection with other disciplines will contribute to understand what appears in the periodicals.

The research question

The objective of the project is, therefore, doing a distant reading analysis of the periodicals included in the repository to better understand how attitudes towards China evolved during the period of the American occupation of the Philippines (1902–1946). From this departing point, the first challenges will be to make the right question(s) so the digital tools are useful in answering it/them. Digitization and the use of digital methodologies allow us to be more ambitious in our research questions. The reflection on the research question and the methodologies that we can use to approach it will provide us with a testing ground for features and tools that we might want to incorporate later to the repository. This can, on the one hand, facilitate the work of researchers in handling both the materials themselves and the metadata. On the other hand, it also enables to set up bridges between librarians and researchers, as well as between Digital Humanities and traditional humanities. According to Huub Wijfjes, this synergy is infrequent, as "advanced digital methods for analysis are not often used to answer concrete research questions in these disciplines", so its usefulness is regularly questioned by humanists (Wijfjes 2017). However, it is not disputed that DH methodologies can be very useful when working on connections in large-scale corpora.

In the case proposed for this chapter, the research will be based on the epistemic problem that Philippine cultural production in Spanish is rarely considered Asian, but rather related to its Western occupiers: North America or Spain. Philippine texts will be approached from three hypotheses: (1) that these texts will evidence the birth and consolidation of a Pan-Asian sentiment, (2) that they will portray the development of a discourse of resistance to Spain and the United States by the identification with China between 1880 and 1930, and (3) that this change was connected to translation and the conception of China in Western countries.

In fact, Teodoro Kálaw, a Filipino writer and politician admitted in the preface to *Sun Yat Sen: Fundador de la República China* by Mariano Ponce, that during the period of Spanish occupation of the Philippines, the attitude of the Filipinos towards the Chinese was one of ignorance and mockery because of the effects of being subjected to a European power (that is, because of the influence of the Spanish) (Ponce 1912, viii, ix). This attitude was supposed to have changed after 1898. Kálaw blamed Spaniards for the negative image of China widespread in the

Philippines, suggesting that it was mediated. At the same time, according to the findings of ALTER group in its 'Archivo China España' Project ('Archivo China España, 1800–1950' n.d.) as stated by Carles Prado-Fonts in "Writing China from the Rest of the West" (2018: 178), knowledge on China in late 19th century Spain often came mediated itself by translations of articles from the United Kingdom, France and to some extent from Germany too.

Therefore, sample the questions to be approached (not answered) here are: (1) are there any similarities or other kind of connections (authorial, ideological…) between the image of China in the Spanish language press and that image in the Philippine press in other languages during the last decade of the 19th century?, and (2) are there any changes in the perception of China along the first decades of the 20th century? These are only proposed as an example of a theoretical research. Translation issues will be crucial in the process as we are addressing a question to be answered through a multilingual set of texts. How can we, with no knowledge of Tagalog or Japanese, track and interpret articles on China in newspapers written in those languages? And how can we connect our findings in those newspapers with what is said in newspapers in English and Spanish? Furthermore, how can we prove if the same article had been translated in different languages and published in different newspapers in the country?

Approaches to studying a country's representation in the periodical press

Amongst the many projects recently set up by historians to profile national identities and their evolution in the periodical press, perhaps one of the most ambitious in methodological terms is the *Trasatlantis* project run by the University of Utrecht ('Transatlantis Locations' n.d.). More than a project, it is a full programme, which began in 2013, on the use of Digital Humanities to talk about "Reference cultures", that is, about "the dominant role of some cultures in the international exchange of ideas, products and practices". Although it is coordinated by the University of Utrecht, the programme involves several Dutch universities and the country's Royal Library, from where the basis of the project, the corpus of periodical publications, is being extracted. They work with exclusively Dutch language texts provided through the *Delpher* ('Delpher – Boeken Kranten Tijdschriften' n.d.) and they have developed a search tool for periodical archives in the library, Texcavator, to obtain statistics, word clouds and timelines, and that allows them to nominalise words and eliminate stopwords ('Texcavator' n.d.). From those, the nominalisation of words is not contemplated in other built-in text mining devices in repositories, such as the Hathi Trust Research Center, a text-mining environment based on Hathi Trust virtual library holdings and developed by the University of Indiana and the University of Illinois at Urbana Champaign ('Our Research Center' 2014; Plale et al. 2013). Although the techniques that they offer are not that complicated and are within reach of anyone who has a corpus, Texcavator and HTRC enable this to be done without downloading the corpus, making it a useful addition to the library's own search engine, which is tailor-made for researchers (Eijnatten, Pieters, and Verheul 2014).

In the case proposed in this chapter, similar text-mining techniques will be applied with not built-in applications to inquire about the extent to which China's image is mediated by Spain or by other powers. Our corpus is an incomplete catalogue of periodical publications initially in image format. At the moment, there are no search tools built in the Philippine repository like Texcavator or like Diacollo, which search for diachronic coincidences around certain key words in the Deutsches Textarchiv (1473–1927) ('D*/DTA Search' n.d.). However, there are several tools that can perform a similar analysis after the corpus is transcribed or OCRd and classified by year and journal.

There are two possible approaches to the problem. One would be working only with headlines. In this case, we can launch a search of Named Entities to check how many articles have China as their main topic and what is said about this country. This can be done in the newspapers in different languages by translating just the named entity if necessary. Working with metadata of the articles will also allow detecting articles by the same authors re-used in newspapers in different languages. For doing so, unless researchers perform an exhaustive hand mark-up of metadata within the newspapers in the repository, proper automated segmentation of areas and recognition of headlines, sub-headings and signatures should be implemented. Google already considered a headlines recognition based on font sizes in 2009 (Chaudhury et al. 2009). In 2012 at the University of Rouen, a method for automatic article detection and indexation was developed and included in their digitization workflow (Palfray et al. 2012, Hébert et al. 2014). In 2015 the method was further developed by considering additional features to font features of headlines, such as word feature, punctuation mark feature, keyword feature and abbreviation feature so they would train the automated identification of headlines in English newspapers (Hanumanthappa and Nagalavi 2015). More recently, the National Library of Finland has developed software to automatically identify and index articles of digitised historical journals (Kettunen, Pääkkönen, and Liukkonen 2019). Going down to the more basic level of an initiation course on Digital Humanities, and without any need of being able to program, the HCR tool Transkribus, which can also be used as OCR, allows to train the text layout and exports the results together with the OCR in ALTO/XML, which already recognises the headlines and marks them. Several groups have acknowledged and worked with the headlines of newspapers, namely, at the University of Leeds, a young group of students has been working with headlines of *The Guardian* for the automatic extraction of six values from the news: prominence, sentiment, superlativeness, proximity, surprise and uniqueness, in order "to facilitate the selection and prioritisation of large volumes of digital content" (Piotrkowicz, Dimitrova, and Markert 2017: 64).

The second option includes considering the whole text for the analysis. In order to do this, periodicals might have to be classified by decades and languages for an observation of the diachronic but also language groups variation in the image of China.

Opinion mining or sentiment analysis can, in semi-absolute terms, help to classify into positive and negative the prevailing attitudes towards China, both inside the country (with reference to the *Sangleys* or Chinese living in the Philippines) and outside it (mentions of "China" as such). The hypothesis put forward by María Dolores Elizalde, who uses traditional methods to analyse various texts by 19th-century travellers and Spanish civil servants, is that for them China has two sides that provoke two different attitudes. On the one hand, China abroad is seen as a great imperial and mysterious nation, which arouses admiration. On the other hand, the Chinese population in the Philippines, according to Elizalde, is connected "with the work they do in the colony" (Elizalde Pérez-Grueso 2008: 101). Therefore, the ideal distant China is the one that prompted idealisations like the ones found in modernist poems of the period, whilst the near China provoked rejection and the terrors of the *yellow peril.* However, Elizalde only considers texts written in Spanish, and we should find out if this differentiation was also present in other languages and in later texts.

Topic modelling can also contribute to ascertaining the motives for this differentiation in attitudes between *Sangleys* and the Chinese outside. It can respond to questions such as (1) what ideas were associated with the *Sangleys*/China, and (2) if they involved the same association of ideas in all languages and in all the regions of the Philippines.

Thus, there are three classic techniques within the wide scope of text-mining that can be suggested for these purposes for beginners in DH: KWIC statistics, sentiment analysis and topic modelling. The two challenges involved are: (1) applying them with no programming involved and (2) using free applications, to ensure sustainability and to cope with the lack of resources that had to be dealt with as the first problem with developing Digital Humanities in the Philippines.

First difficulty: How to prepare a set of plurilingual texts?

Until very recently, Philippine history studies on the Philippines used *The Philippine Islands* collection by Blair and Robertson as primary sources, a series of Filipino archive documents translated from Spanish into English in 55 volumes (Cano 2008: 236). The collection of volumes was part of the North American colonial project, meaning that using only these materials prevented a complete and objective study of the past in the Philippines from being undertaken until 1898 (the date on which the volumes on *The Philippine Islands* arrived). This practice illustrates the need to be able to cover documents in all the languages they are available in, thereby resolving researchers' lack of knowledge of all these languages. Each language represents a political and conceptual view of the world and of Filipino society that is systematically forgotten in studies on the archipelago's monolingual texts, meaning that one of the fundamental aims of Digital Humanities should be directed at settling this situation.

Although major advances in *deep learning* have led to Google, Microsoft and other major automated translation systems to using neuronal networks and substantially improving their translations in what is called Neural Machine Translation (Castelvecchi 2016), applying it to the periodicals corpus we are working on results in three problems:

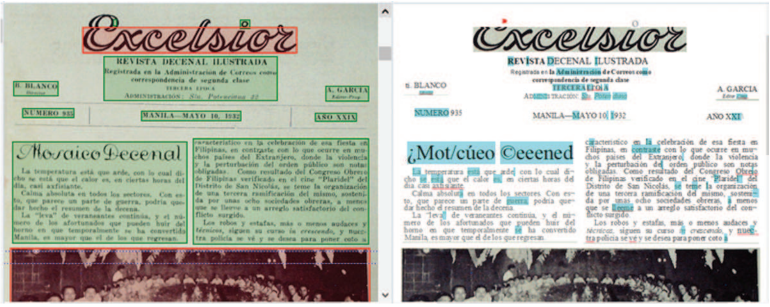
1. The quality of the texts produced from images (OCR) is not enough, with too many errors in the original language that are multiplied when the texts are translated.
2. The text layout on periodical pages causes confusion in the OCR, which, depending on the layout analysis performed, may read lines in different columns as continuous lines.
3. We are working with low-resource languages, with insufficient digital input and barely any tools developed for them.

Before addressing these problems, we are attending briefly to the platform problem. In order to do the repository that we are to work with sustainable and flexible enough to have the possibility to add new utilities if necessary, the VLIRUOS-TEAM project is using Universal Viewer. This visualiser of documents supports the IIIF manifesto and is making great progress in viewing, sharing and searching in periodicals ('IIIF Newspapers Community Group – IIIF | International Image Interoperability Framework' n.d.). Universal Viewer is the viewer used by organisations like the press collection at the National Library of Wales ('Welsh Newspapers Online – Home' n.d.), which published the code created to facilitate the inclusion of transcribed text by articles alongside the image and the search for text in the complete collection for replication in other repositories ('IIIF Newspapers – Devwiki' n.d.). Progress is being done also to integrate OCRd text together with the IIIF image view, and to allow cross-searches of words across the repository.

The problem with OCR

The OCR problems are normally enumerated as widespread challenges to be taken into account in historical periodical repositories. OCR was the main topic of issue number 13 of EuropeanaTech ('OCR' 2019). To overcome OCR issues, they point at the improvements developed thanks to engines that utilise deep neural networks such as Tesseract, OCRopus, Kraken and Calamari (*Tesseract-Ocr/Tesseract* [2014] 2020; Tom [2014] 2020; *Calamari-OCR/Calamari* [2018] 2020; 'Kraken – Kraken 2.0.5-4-Gbb42ba5 Documentation' n.d.). Our situation here is, however, that there is a lack of expertise in programming, necessary for the implementation of this software.

Although document scanners usually have an OCR software integrated, like ABBYY, which includes a layout recogniser and can produce ALTO/XML, the poor quality of some of the materials confirms the already well-known error rate of this OCR. ABBYY is moreover language bond. It means that their predictions are based on a number of languages pre-set. It thereby excludes -or posses a difficulty- to OCR the texts in minority languages present in our corpus, such as Cebuano or Ilokano. Even in texts in majority languages like Spanish, the mix of fonts and of text with image and artistic text produces area and OCR detection errors (see Figure 3).



Excelsior
REVISTA DECENAL ILUSTRADA
Registrada en la Administracion de Correos como correspondencia de segunda clase
TERCERA EPOCA
ADMINISTRACIÓN: Sto. Potenciana 22
B. BLANCO
Director
A. GARCIA
Editor-Prop
NUMERO 935
MANILA—MAYO 10, 1932
AÑO XXIX

Mosaico Decenal

La temperatura está que arde, con lo cual dicho se está que el calor es, en ciertas horas del día, casi axfisiante.

Calma absoluta en todos los sectores. Con esto, que parece un parte de guerra, podría quedar hecho el resumen de la decena.

La "leva" de veraneantes continúa, y el número de los afortunados que pueden huir del horno en que temporalmente se ha convertido Manila, es mayor que el de los que regresan.

característico en la celebración de esa fiesta en Filipinas, en contraste con lo que ocurre en muchos países del Extranjero, donde la violencia y la perturbación del orden público son notas obligadas. Como resultado del Congreso Obrero de Filipinas verificado en el cine "Plaridel" del Distrito de San Nicolás, se teme la organización de una tercera ramificación del mismo, sostenida por unas ocho sociedades obreras, a menos que se lleve a un arreglo satisfactorio del conflicto surgido.

Los robos y estafas, más o menos audaces y *técnicos*, siguen su curso *in crescendo*, y nuestra policía se vé y se desea para poner coto a

REVISTA DECENAL ILUSTRADA
Registrada en la Administración de Correos como correspondencia de segunda clase
TERCERA EPOCA
ADMINISTRACIÓN: Sto. Potenciana
B. BLANCO
A. GARCIA
NUMERO 935
MANILA—MAYO 10, 1932
AÑO XXI

¿Mot/cúeo ©eeened

La temperatura está que arde con lo cual dicho se está que el calor es, en ciertas horas del dia casi axfisiante.

Calma absoluta en todos los sectores. Con esto, que parece un parte de guerra podría quedar hecho el resumen de la decena.

La "leva" de veraneantes continúa, y el número de los afortunados que pueden huir del homo en que temporalmente se ha convertido Manila, es mayor que el de los que regresan.

característico en la celebración de esa fiesta en Filipinas, en contraste con lo que ocurre en muchos países del Extranjero, donde la violencia y la perturbación del orden publico son notas obligadas. Como resultado del Congreso Obrero de Filipinas verificado en el cine "Plaridel" del Distrito de San Nicolás, se teme la organización de una tercera ramificación del mismo, sostenida por unas ocho sociedades obreras, a menos que se lleeme a un arreglo satisfactorio del conflicto surgido.

Los robos y estafas, más o menos audaces y *técnicos*, siguen su curso *in crescendo*, y nuestra policía se vé y se desea para poner coto a

Figure 3. Recognition of text areas with ABBYY FineReader 14 with numerous errors, both in the text and in the recognition of areas of image and text in the 10 May 1912 issue of Excelsior magazine, scanned at the recommended resolution of 500pp

The old solution of double-keying, which used to be popular for data entry, has indeed a high accuracy rate (Haaf, Wiegand, and Geyken 2013). This involves two people manually entering text extracted from an image and comparing the two versions. However, due to the high budget required for this operation, efforts have turned to find more time and money efficient methods to extract text. Another explored solution which has been popular is crowdsourcing (Hedges and Dunn 2017: 351–52). The National Library of Australia, on its historical periodicals site, Trove, combines automated OCR with crowdsourcing for correcting errors, which seems to be an acceptable option, although the volunteer work process is necessarily slow ("Trove – Digitised Newspapers and More" n. d.).[4] However, in the sphere of cooperation this possibility creates ethical conflict: free work in developing countries, even if it is non-profit, does not seem acceptable. Beyond this, crowdsourcing also calls into question the quality of the result and the possibility of covering the entire corpus.

A possible solution is Transkribus. This is an opensource desktop application for the automated transcription of historical documents which constitutes the basis for the READ project (Recognition and Enrichment of Archival Documents). It was implemented by the European Association for Digital Humanities (EADH) and aspires to "set new standards in Handwritten Text Recognition, Key Word Spotting, Layout Analysis, Automatic Writer Identification and related fields" ('READ | EADH – The European Association for Digital Humanities' n.d.). The platform, initially conceived for Handwriting recognition, allows to perform an automated layout analysis and OCR (based on ABBYY for printed texts) and to train a model by correcting and saving some pages. Whilst for HRC the minimum number of pages is around 150, for printed documents it works with a high accuracy rate with fewer pages trained. In fact, the Impresso project run by the Zurich Computational Linguistics Group presented a poster in November 2018 at the DARIAH-CH workshop in which they used Transkribus for character recognition in historical periodicals with positive results: firstly, they manually transcribed 150 random front covers of different newspapers from different centuries (that is, with different fonts) and trained the model in Transkribus. They managed to get the error ratio down to 2%, resulting in Transkribus ('Transkribus' n.d.) being accepted as a valid solution for extracting text from historical periodicals on a grand scale (impresso 2018). In 2019, Günter Mühlberg and Günter Hackl published a training dataset of 148 pages of Austrian newspapers and a validation set of 13 pages, which endured even further facilitation for the researcher (Guenter Muehlberger and Guenter Hackl 2019). In Utrecht DH2019, the group from Zurich proved that besides the initial success with the OCR, the HRC was still very useful for newspapers printed with black letter, with a 97% rate of accuracy (Ströbel and Clematide 2019), which is especially useful for extracting text from damaged images such as microfilms.

[4] Since June 2008 (when corrections started) through to January 2019, a total of 296,268,735 lines have been corrected thanks to 53,978 registered correctors. As at January 2019, a total of 24,875,005 digitised pages are in the repository ('About Newspapers' n.d.; 'Text Correction Hall of Fame' n.d.).

Translation

In the hypothetical workflow in which we are working, after digitising and uploading images and text to the visualiser (images to Universal Viewer and text as commentaries), the next step should be to carry out some analysis tasks. At this point is when we need to confront two problems: the multilingualism of our corpus and the fact that some of the languages used in it are low-resourced languages. It means that there are no large-scale parallel corpora available.

Given this situation, we find two options. On the one hand, working with systems that allow multilingual text-mining (and therefore there is no need for translation), and on the other hand, attempting an automated translation of the corpus. Multilingual text-mining will be approached in the next section. Here we are focusing on the possibility of performing an automated translation of the corpus.

Machine Translation (MT) of low-resourced languages is a rapidly developing field in Computer Linguistics and Natural Language Processing. Big companies like Facebook are developing ways to overcome this difficulty and they are implementing a combination of methods for Unsupervised Machine Translation for low-resourced languages. They include word-by-word initialization, iterative back-translation and language modelling ('Unsupervised MT: Fast and Accurate for More Languages' 2018). They are also open-sourcing the tools that they have created ('LASER NLP Toolkit: Zero-Shot Transfer across 93 Languages' 2019). There are at the moment at least two special issues on Machine Translation for Low-Resource Languages being prepared simultaneously in 2020: one for the journal *Machine Translation* and another one for the *American Journal of Artificial Intelligence*. Meanwhile, the topic continues being prominent in NLP conferences (Gu et al. 2018; Pa et al. 2016). All these promising developments are however difficult to implement for a project to be run by beginners in DH with no programming knowledge. However, a painstakingly easy provisional option could be adding a Google Translate plugin to Omeka. Currently, Google Translate allows translations to and from Tagalog and Cebuano, however it does not include translation into other Filipino languages like Ilokano, also present in our corpus. It is true that Google has only been in the Philippines for three years and a number of messages have already appeared in the forums asking it to include Ilonggo/Hiligaynon and Ilokano to its translation tool. Moreover, the possibility of translating Filipino periodicals from Spanish, English, Japanese, Tagalog and Cebuano covers 98% of the corpus.

Making these Filipino languages visible in the DH community is however important to demand online tools and resources for them and to contribute to moderate the prevalence of English in the discipline (Crompton, Lane, and Siemens 2016: 27).

Tools

For working with just one language, there is a wide range of digital tools to process the text. Voyant Tools or Antconc for example can do a fantastic job with word clustering, and with simple statistical analysis such as cooccurrences, keywords and visualizations similar to the ones obtained with more complete statistical packages like 'R'.

Listed below are some of the possible resources for contributing to finding an answer to the research question and their specific usefulness for this theoretical case study:

Voyant-Tools:[5] this tool can be used to explore sets of text documents (i.e. corpus of one periodical title or several by year), as well as word frequency, for example, the coincidence of terms grouped together by the researcher. That is, by grouping the collection of a single title into several years, we can see, for example, in which years/months there are more mentions of the Chinese – which can be compared with historic events taking place at that time – or of several titles, in which words like "Chin[a|ese]", "empir[e|s] / imperial", "silk", "pigtail" and "dirty" coincide. Similar collocations and antagonism can be located in periodicals in English. Does Chin[a|ese] appear in the same texts in which "Japan[ese]" and "Asia[n]" and "Filipin[o|os]" appear? These searches will enable us to decode references to the Chinese inside the Philippines, the Chinese as an Asian country and some of the characteristics assigned to them, depending on the language of the periodicals and the moment in history when they were published (for example, around the time of the war between China and Japan). The choice of search terms must be based on an earlier search or on previously formulated suppositions and hypotheses. One important problem that this tool poses, however, is that it supports a very limited corpus in terms of weight.

Antconc:[6] this popular tool can be used for obtaining results with a lower intuitive percentage of collocations around the idea of China, for example. In Antconc you can work with text versions of periodicals or with linguistically labelled versions (POS), for example with Freeling or with TagAnt, by the same creator of Antconc.

[5] https://voyant-tools.org/

[6] http://www.laurenceanthony.net/software/antconc/

> This means that collocations around a word or keywords can be located (for example, Chin[a|ese]) by searching for concordances (KWIC), eliminating stop words or perhaps, if the text is syntactically labelled, by selecting only nouns and adjectives around the chosen words. Once this has been done, it is not difficult to copy the words around the keyword (China) and view the most frequent ones in the selected corpus in a word cloud, divided by language or perhaps by year. From there you can also look at how terms are scattered to see where some coincide more, as done with Voyant-tools. An example of recent successful use of Antconc combined with Wordcloud for visualization is the paper "The Media Construction of Italian Identity", which explores newspapers written in Italian in Italy and in the US to examine *la italianità* as an identity discourse between the end of 19th and the beginning of 20th century (Viola and Verheul 2019). The idea that the Chinese presence in the Philippines challenges the Hispanicity of the country as well as the Spanish interests in Asia is present in several Spanish books and reports of the end of the 19th century. A similar methodology to the one used in the mentioned article might produce interesting ideas around the questions approached in this paper.

Despite needing some knowledge on its programming language, Matthew Jockers has a very useful and easy to follow guide to *Text Analysis with R for Students of Literature* (Jockers 2014: 47). Jockers explains the steps to perform similar experiments like the ones described for Antconc and Voyant-Tools. He adds the possibility of text modelling with the *Mallet* package for R, which cannot be done for the moment with the former two tools. It has the additional feature of viewing the correlation of words around two key terms, for example, China – Philippines. Topic modelling may contribute to identifying recurring themes and terms related to those themes in the corpus inserted. Therefore, we can see what is being said about the Chinese in general when discussing the Chinese and whether more than one theme is related to them (for example, money, jewellery, furniture, trade or dirt). Table 1 shows the results of performing *topic modelling* with Mallet in R of Spanish books about the Chinese in the Philippines written in 1884, 1886, 1888 and 1892, in which sub-themes can be seen in relation to the Chinese (Jordana y Morera 1888; *La Inmigración China y Japonesa En Filipinas: Documentos* 1892; Comenge 1894; *Los Chinos En Filipinas: Males Que Se Experimentan Actualmente y Peligros de Esa Creciente Inmigración* 1886). Based on this modelling readings can be made in which a combination of words is selected (for example "chino" "sangley") to see what words or themes are grouped around these two concepts. Although using Mallet in R may seem to be more complicated than otherwise, Matthew Jockers provides excellent instructions for its use and code that can be reused, making the operation a much easier one (Jockers 2014: 135–52).

Table 1. List of topics of four books written by Spaniards on the Chinese presence in the Philippines. On the left are the groupings performed by Mallet, on the right are the names given to the topic according to words grouped by Mallet

	Most frequent words of the topics	Name of the topic
1	gobernador españoles sangleyes manila chinos ciudad	Situation of the Chinese in Manila
2	china rey nombre mar gente tierra oro	Wealth of the country of China
3	chinos comercio manila chino país china años	Chinese merchants resident in Manila
4	fuerte ejército salcedo sioco campo indios limahon	Armed conflict
5	filipinas archipiélago islas gobierno chino país china	Territories of China and the Philippines
6	isla clase tienda tiendas azúcar abacá efectos	Trade in the Philippines (products)
7	capitana gente juan galera enemigo noche oidor	Night boat danger
8	altar china misioneros cristianos santa padre ritos	Christian missionaries in China
9	sangleyes indios real tierra indias ley gobernador	Legislation related to the Sangleys
10	confucio zu imperio letra emperador cielo libros	Imperial China high culture
11	chino té seda casa papel cantón arroz	Chinese culture
12	fiesta ídolo día sol año luna	Chinese festivals
13	negocios provincias personas banco administración crédito públicos	Economics
14	raza europea isla defender sociedades norte blanca	European defenders of the Philippines
15	inmigración oceania periódico número seguridad medios	La Oceanía newspaper
16	chinos general decreto cédula hacienda sangleyes reglamento	Legislation Chinese in the Philippines
17	noche opio juego ladrón cierta extremo	Vices of the Chinese
18	opio renta pesos fumaderos mil estado	Smoking dens
19	tao pueblo virtud sabio sabe llama	Chinese wisdom
20	chino hijos matrimonio sangley padres familia cristiano	Chinese families

The last technique of the ones considered to be useful for examining this large quantity of periodicals is sentiment analysis. Programming Historian has an excellent introduction to the concept and the use of sentiment analysis using the Natural Language Toolkit with Python (Saldaña 2018). However, the code they provide is for working with texts in English. Something similar happens with other tools: xTas is a wrapper with a range of tools for text work, including one for sentiment analysis, but it works with Dutch and English ('Xtas, the EXtensible Text Analysis Suite – Xtas 3.4 Documentation' n.d.). The "Mining Shifting Concepts through Time" (ShiCo) project is developing "a tool that enables humanities researchers to mine the historical development of concepts" ('Netherlands EScience Center' n.d.). It is based on "vector representations derived by neural network language models". The software they have created works with Word2vec, so corpus preparation may be a little more complex than with other programs. It is not exactly a sentiment analysis program but can track concepts over time and connect them together with graph views of how certain concepts are related to others at a particular time in order to see how they evolve (*Netherlands EScience Center: Shifting Concepts Through Time Project – NLeSC/ShiCo* [2015] 2018).

The problem of trying to apply sentiment analysis in our corpus is precisely that in the majority of cases the vocabularies used on which the analysis is based are in English (or in Dutch). It is not difficult to find vocabularies in Spanish for this purpose; José Cardona Figueroa has published one in Github (Figueroa [2015] 2018), but for languages like Ilokano or even Tagalog, it is a much more difficult matter. In such cases, one solution can be to create your own thesaurus. Another solution might be translating the corpus in order to perform the analysis. Sentiment analysis simplifies the themes explored previously and reveals if most of the comments on one notion (in this case China) are positive or negative, and in what sense (fear, disgust, anxiety, happiness, etc.) in a few categories set by the thesaurus. By using ShiCo it is possible to observe the evolution over time of ideas around the key concept. The concept relation graphs add, moreover, a diachronic component: you can see something that looks similar to what we saw in topic modelling as it moves forward in time.

There is a second option for this operation which is multilingual text analysis. Since the 1990s multiple groups have explored the possibilities of performing distant reading operations and information retrieval on multilingual sets of texts. Most recently, in 2019, a book, *Multilingual Text Analysis, Challenges, Models and Approaches* has been published by Natalia Vanetik and Marina Litvak (2019) that provides a holistic approach to the subject and the previous and current efforts in the field. Noteworthy is the abovementioned Newseye project, which combines some of the previous experiences to apply it to a multilingual corpus of newspapers. It is indeed one of the declared aims of the project: providing solutions for text recognition problems in multilingual semantic text enrichment and in dynamic text analysis ('Aims' n.d.). Applied to topic modelling, they have created a method to combine Dynamic Topic Modelling (DTM), that takes into account the evolution of topics over time, with Multilingual Topic Modelling (MLTM), which, by contrast with using the most common LDA (Latent Dirichlet Allocation) for each language, allows alienation of topics across languages ('On Multilingual Dynamic Topic Modeling' n.d.). In this way, by the combination of both in what they have called Multilingual Dynamic Topic Modeling (ML-DTM), they have managed to identify events around a topic aligned in a cross-lingual manner (Zosa and Granroth-Wilding 2019). This is indeed promising in order to run analysis on the perception of China in parallel comparable sets of at least bilingual newspapers along time. In this sense, we could provide partial answers to both of the central questions that are to be addressed in this paper: due to the multilingual topic modelling connections between the image of China in the Spanish language press and in other languages from the Philippines could be compared, in terms of seeing in parallel which topics are connected to China. The diachronic question of the evolution of these perceptions between the end of 19th and the beginning of the 20th century could be also addressed thanks to the Dynamic Topic Modelling.

A further possibility would be weighing the results with what is happening in other parts of the world by establishing comparisons with sets of newspapers held in other digital repositories to check whether ideas being discussed in the Philippines are related to those published in other countries like Spain, the United States, France and the United Kingdom. The first two played a dominant role in sectors of Filipino society in the first half of the 20th century as former coloniser and new coloniser. Part of society, usually the dominant classes, were in favour of the country following Spain's lead and forming part of the community of Hispanic countries in Latin America in a discourse that became known as *Hispanidad*, which idealised Spain's role in the history of its former colonies and advocated it as a maternal, protective figure for the new republics. Whereas the new generation of Manila argued for the modernisation put forward by the Americans and was in favour of the ideas coming in from the United States, scorning the nostalgia for everything Spanish as outdated. Meanwhile, England and France were influential centres of creativity and intellectualism across the entire world between the second half of the 19th century and the first half of the 20th century.

Some various open repositories and aggregators allow work with the periodicals in one or several countries. On its library website, Cornell University provides a list of online repositories of both current and historical periodicals across the world, which can be very useful.[7] The CLARIN European network also pro-

[7] https://guides.library.cornell.edu/news_online

vides users with lists of periodical repositories, some with extra tools like online concordance search engines.[8] Most of the corpora listed on their website can be downloaded too. One of the most popular aggregators in Europe on Europeana,[9] which has recently been developing its newspaper branch with excellent advice on how to manage metadata in repositories to make the researcher's job an easier one. Europeana also allows you to search and work with the full text of collections from several countries in different languages (Willems and Atanassova 2015).

The Multilingual Dynamic Text Analysis, however, involves an advance knowledge in programming. Although it is an old discussion the matter of the capability of Digital Humanists to program, with the initial circumstances that we face and the objective of spreading DH among a community of humanists, the challenge would be creating an integrated tool to perform this kind of text analysis within the repository.

What would researchers in the humanities need from a periodicals repository in the 21st century?

In this hypothetical journey through the various steps and tools available for resolving a research problem with a corpus of periodicals, we have had to handle the corpus in different ways. For this handling, the way periodicals are presented in the repository can make an enormous difference in terms of making the researcher's job easier. Nowadays, there are several initiatives in the Philippines for building virtual repositories for books and periodicals,[10] led by the prodigious and pioneering work of the *Digital Library* at the University of Santo Tomás in Manila which, funded by UnionBank, has already digitised 63 titles of *rare periodicals*. The repository, although it fulfils a priceless job of preserving and disseminating the material, is very difficult to use from the digital humanities perspective: the materials cannot be downloaded, they are only available in a non-readable, low-quality format, so there is no possibility of doing text searches across the whole corpus (as the text has not been extracted), it is built on a paying platform that compromises

[8] https://www.clarin.eu/resource-families/newspaper-corpora

[9] http://www.europeana-newspapers.eu/

[10] Philippine Heritage Library is uploading digitised non-copyrighted books to its own catalogue and the National Library of the Philippines has a plan for doing the same. López Museum and Ortigas Foundation are also digitising their collections. For a landscape on digitisation in the Philippines see Ortuño Casanova and Sarmiento 2020.

its long-term survival, and the metadata used are not standardised nor can they be downloaded, nor has an OLR been done.[11]

Nowadays, researchers, librarians, programmers and curators are finally holding discussions on how the effort and money invested in creating repositories might be optimised. A widespread desire expressed in multiple venues -and most lately by the Ocean Exchanges panels in DH Utrecht 2019, was the necessity for common standards and methodologies to address common challenges, noting text reuse in newspapers across countries as one of these challenges (Cordell n.d.).

Some organisations have published their advice on metadata standardisation and even their own code for modifying viewers with the open intention that other periodical repositories imitate them. This is the case of the above-mentioned historical newspaper repository in the National Library of Wales, which has a Devwiki with the clear objective that it will "hopefully be replicated by other IIIF institutions so that code developed against the Welsh Newspapers will also work with other sites" ("IIIF Newspapers – Devwiki" n. d.). For years now, projects have been publishing good practice manuals which should now be gradually updated. As an example, the Neptuno Project manual dates back to 2004 (Castells et al. 2004) and in 2016 the heads of Europeana Newspapers presented a kind of wish list of the implementations they were going to put in place in their aggregator to create a common search engine for several repositories in an article intriguingly entitled "Making Europe's Historical Newspapers Searchable" (Neudecker and Antonacopoulos 2016). Partnerships between private companies and institutions are also being set to standardise and facilitate the whole digitization flow. This is the case of Kitodo, a partnership between "open source community and service providers" launched by the Hamburg State and University Library, that supports the digitization of cultural products from digitization itself to construction of metadata and online publication ('Issue 10: Innovation Agenda' n.d.).

Regarding the connection between librarians, developers and researchers using bibliographic information, at the end of 2019 a new DARIAH working group was set up with these specific aims: Bibliodata ('Bibliographical Data (BiblioData) | DARIAH' n.d.). In the Philippine case, a forum connected to the VLIRUOS-TEAM project in which the Universiteit Antwerpen and the University of the Philippines are partners was held in November 2019 to enhance communication between researchers in the humanities and librarians, focused on digitization processes.

However, in order to process the texts, once they have been digitised and OCRd, keeps on being challenging for Humanities Researchers. What we have proposed here are basically operations that are typical of Natural Language Processing and

[11] http://digilib.ust.edu.ph/rare-perio.html

Computer Linguistics applied to newspapers (including literature and news) in order to respond to a humanities question on representation. Although interdisciplinarity is highly encouraged by institutions, as well as humanists are encouraged to acquire digital skills, the methodologies used are from quite a different area and therefore it is complex to master for scholars from outside that area. In this sense, repositories, librarians and programmers have the very important task of bridging that gap by incorporating tools to repositories that allow work with the corpus that they hold. This may be done either by suggesting self-standing tools that can be used with that corpus or by integrating the text analysis tools in the repository itself in order to produce data out of the corpus. In the proposed case, to resolve the question being discussed of how the image of the Chinese in the Philippines varies from linguistic group to linguistic group from and over time, the following implementations in the repository would be very welcome:

1. Well-transcribed texts (one possibility would be with Transkribus) that can be seen and downloaded in text format, as well as a simplified viewing option in the repository (bearing in mind that the Philippines has one of the slowest internet connections in the world and that the servers will be located there, a simplified version would speed consultations up considerably).
2. Metadata in a standardised format and compatible with formats in other repositories to make aggregating and viewing periodicals easier, as well as downloading the metadata either manually or using programs that allow subsequent work to be done with them (ie. RIS, which works with open bibliography managers as Zotero).
3. Periodicals available for browsing by date, title, languages and places, as well as by searching for keywords in both the text and in the metadata (including titles and authors of articles).
4. Repositories with IIIF implemented and using a compatible viewer.
5. Both the viewer and the platform made free of charge and with a community to maintain them as well as backing to prevent the repository from being discontinued when funding runs out.
6. Standard markup and labelling common across other repositories.
7. A translation tool supplied as part of the repository itself.
8. Downloading high-quality images as one of the options provided.
9. Integration or suggestion of text processing tools and data on the repository to facilitate the humanities researchers' work.

If previous approaches to the image of the Chinese in the Philippines have had no other option than including close reading of a monolingual reduced corpus, the implementation of the abovementioned suggestions may provide credible generalisations. Easy to reproduce, agreeing with standards of deontology in disciplines working with quantitative methods, and more relevant for the development of knowledge in the sense that the output, a landscape of the opinions and their evolution, can be actually put in context and explain social, artistic, economic and literary processes. This cannot be done with close reading, which imposes a very limited selection of texts.

References

"About Newspapers". n.d. Trove. Accessed 25 January 2019. https://trove.nla.gov.au/newspaper/about

"Aims". n.d. Accessed 25 January 2019. https://www.newseye.eu/project/aims/

"Antwerp Centre for Digital Humanities and Literary Criticism – ACDC – University of Antwerp". n.d. Accessed 25 January 2019. https://www.uantwerpen.be/en/research-groups/digitalhumanities/

"Archivo China España, 1800–1950". n.d. Accessed 4 November 2018. http://ace.uoc.edu/

Benson, Rodney, and Erik Neveu. 2005. "Introduction: Field Theory as a Work in Progress". In *Bourdieu and the Journalistic Field*, 1–24. Cambridge, UK: Polity Press.

"Bibliographical Data (BiblioData) | DARIAH". n.d. Accessed 2 February 2020. https://www.dariah.eu/activities/working-groups/bibliographical-data-bibliodata/

Calamari-OCR/Calamari. (2018) 2020. Python. Calamari-OCR. https://github.com/Calamari-OCR/calamari

Cano, Glòria. 2008. *De Tartessos a Manila: Siete estudios coloniales y poscoloniales*. Edición: 1. València: Publicacions de la Universitat de València.

Castells, P., F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras, and J. Lorés. 2004. "Neptuno: Semantic Web Technologies for a Digital Newspaper Archive". In *The Semantic Web: Research and Applications*, edited by Christoph J. Bussler, John Davies, Dieter Fensel, and Rudi Studer, 445–58. Lecture Notes in Computer Science. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-25956-5_31

Castelvecchi, Davide. 2016. "Deep Learning Boosts Google Translate Tool". *Nature News*. https://doi.org/10.1038/nature.2016.20696

Chaudhury, K., A. Jain, S. Thirthala, V. Sahasranaman, S. Saxena, and S. Mahalingam. 2009. "Google Newspaper Search Amp;#150; Image Processing and Analysis Pipeline". In *2009 10th International Conference on Document Analysis and Recognition*, 621–25. https://doi.org/10.1109/ICDAR.2009.272

Comenge, Rafael. 1894. *Cuestiones filipinas. 1a. parte. Los Chinos. (Estudio social y político)*. Manila: Tipolitografía de Chofré y compañía.

Cordell, Ryan. n.d. "Our Project Team". Accessed 2 February 2020. http://oceanicexchanges.github.io/team/

Crompton, Constance, Richard J. Lane, and Ray Siemens. 2016. *Doing Digital Humanities: Practice, Training, Research*. Taylor & Francis. https://doi.org/10.4324/9781315707860

"D*/DTA Search". n.d. Accessed 25 January 2019. http://kaskade.dwds.de/dstar/dta/

"Delpher – Boeken Kranten Tijdschriften". n.d. Accessed 25 January 2019. https://www.delpher.nl/

Eijnatten, Joris van, Toine Pieters, and Jaap Verheul. 2014. "Using Texcavator to Map Public Discourse". *Tijdschrift Voor Tijdschriftstudies*, July, 59–65. https://doi.org/10.18352/ts.303

Elizalde Pérez-Grueso, María Dolores. 2008. "China – España – Filipinas: percepciones españolas de China – y de los chinos – en el siglo XIX". *Huarte de San Juan. Geografía e historia*, no. 15: 101–11. dialnet.unirioja.es/servlet/articulo?codigo=3074412

Figueroa, José Cardona. (2015) 2018. *Contribute to JoseCardonaFigueroa/Sentiment-Analysis-Spanish Development by Creating an Account on GitHub*. R. https://github.com/JoseCardonaFigueroa/sentiment-analysis-spanish

"Fire Breaks out at UP Diliman Campus". 2016. Cnn. 2016. http://cnnphilippines.com/metro/2016/04/01/up-diliman-faculty-center-fire.html

"Fire Hits National Archives Building". 2018. Philstar.Com. 28 May 2018. https://www.philstar.com/headlines/2018/05/28/1819408/fire-hits-national-archives-building

GMA News Online. 2016. "Namria Discovers 400 to 500 New Islands in PHL Archipelago", 2016. http://www.gmanetwork.com/news/story/555068/news/nation/namria-discovers-400-to-500-new-islands-in-phl-archipelago/

Gu, Jiatao, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. "Universal Neural Machine Translation for Extremely Low Resource Languages". In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 344–354. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1032

Guenter, Muehlberger, and Guenter Hackl. 2019. "NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.)". Zenodo. https://doi.org/10.5281/zenodo.3387369

Haaf, Susanne, Frank Wiegand, and Alexander Geyken. 2013. "Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text". *Journal of the Text Encoding Initiative*, no. Issue 4 (March). doi: https://doi.org/10.4000/jtei.739

Hanumanthappa, M., and Deepa Nagalavi. 2015. "Identification and Extraction of Headlines from Online English Newspaper- Statistical Approach" 10 (January): 19–22.

Hébert, David, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez, and Thierry Paquet. 2014. "Automatic article extraction in old newspapers digitized collections". In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*. Association for Computing Machinery, New York, 3–8. https://doi.org/10.1145/2595188.2595195

Hedges, Mark, and Stuart Dunn. 2017. *Academic Crowdsourcing in the Humanities: Crowds, Communities and Co-Production*. Chandos Publishing.

"IIIF Newspapers – Devwiki". n.d. Accessed 25 January 2019. https://dev.llgc.org.uk/wiki/index.php?title=IIIF_Newspapers

"IIIF Newspapers Community Group – IIIF | International Image Interoperability Framework". n.d. Accessed 25 January 2019. https://iiif.io/community/groups/newspapers/

Impresso. 2018. "Moving beyond Digital Filters. How to Integrate the Digitised Press into the Historian's Workflow". Blogpost. Impresso. 6 July 2018. https://impresso-project.ch/news/2018/07/06/laurel.html

"Issue 10: Innovation Agenda". n.d. Europeana Pro. Accessed 3 February 2020. https://pro.europeana.eu/page/issue-10-innovation-agenda

Jockers, Matthew Lee. 2014. *Text Analysis with R for Students of Literature*. https://doi.org/10.1007/978-3-319-03164-4

Jordana y Morera, Ramón. 1888. *La inmigración china en Filipinas*. Madrid: Tipografía de Manuel G. Hernández.

Kettunen, Kimmo, Tuula Pääkkönen, and Erno Liukkonen. 2019. *Clipping the Page -Automatic Article Detection and Marking Software in Production of Newspaper Clippings of a Digitized Historical Journalistic Collection*. https://doi.org/10.1007/978-3-030-30760-8_33

"Kraken – Kraken 2.0.5-4-Gbb42ba5 Documentation". n.d. Accessed 1 February 2020. http://kraken.re/

La Inmigración China y Japonesa En Filipinas: Documentos. 1892. Madrid: Imprenta de Don Luis Aguado.

Lagrama, Eimee Rhea C. 2012. "Preventing Disaster: Quantifying Risks at the UP Diliman University Library". In *Libraries, Archives and Museums: Common Challenges, Unique Approaches*, 10. Rizal Library. Ateneo de Manila University.

"LASER NLP Toolkit: Zero-Shot Transfer across 93 Languages". 2019. 22 January 2019. https://ai.facebook.com/blog/laser-multilingual-sentence-embeddings/

Li, David Leiwei. 2003. *Globalization and the Humanities*. Hong Kong University Press.

Los chinos en Filipinas: Males que se experimentan actualmente y peligros de esa creciente inmigración. 1886. Manila: Establecimiento tipográfico de La Oceanía Española.

"Netherlands EScience Center". n.d. Accessed 29 January 2019. https://www.esciencecenter.nl/project/mining-shifting-concepts-through-time-shico

Netherlands EScience Center: Shifting Concepts Through Time Project – NLeSC/ShiCo. (2015) 2018. *Python*. Netherlands eScience Center. https://github.com/NLeSC/ShiCo

Neudecker, C., and A. Antonacopoulos. 2016. "Making Europe's Historical Newspapers Searchable". In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 405–10. https://doi.org/10.1109/DAS.2016.83

"OCR". 2019. 13. EuropeanaTech. Europeana. https://pro.europeana.eu/page/issue-13-ocr

"On Multilingual Dynamic Topic Modeling". n.d. Accessed 2 February 2020. https://www.newseye.eu/blog/news/multilingual-dynamic-topic-modelling/

Ortuño, Casanova Rocío. 2017. "Philippine Literature in Spanish: Canon Away from Canon". *Iberoromania* 2017 (85): 58–77. https://doi.org/10.1515/iber-2017-0003

Ortuño Casanova, Rocío and Anna Sarmiento. 2020. "Humanidades Digitales en Filipinas: proyectos, dificultades y oportunidades de la colaboración Norte-Sur". *Digital Scholarship in the Humanities, fqz086*. https://doi.org/10.1093/llc/fqz086

"Our Research Center". 2014. HathiTrust Digital Library. 2014. https://www.hathitrust.org/htrc

Pa, Win Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. 2016. "A Study of Statistical Machine Translation Methods for Under Resourced Languages". *Procedia Computer Science*, SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09–12 May 2016 Yogyakarta, Indonesia, 81 (January): 250–57. https://doi.org/10.1016/j.procs.2016.04.057

Palfray, Thomas, David Hebert, Stéphane Nicolas, Pierrick Tranouez, and Thierry Paquet. 2012. "Logical segmentation for article extraction in digitized old newspapers". In *Proceedings of the 2012 ACM symposium on Document engineering (DocEng '12)*. Association for Computing Machinery, New York, 129–132. https://doi.org/10.1145/2361354.2361383

"Philippines". n.d. Ethnologue. Accessed 18 September 2018. https://www.ethnologue.com/country/PH

Piotrkowicz, Alicja, Vania Dimitrova, and Katja Markert. 2017. "Automatic Extraction of News Values from Headline Text". In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 64–74. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/E17-4007. https://doi.org/10.18653/v1/E17-4007

Plale, Beth, Robert McDonald, Yiming Sun, Inna Kouper, Ryan Cobine, J. Stephen Downie, Beth Sandore Namachchivaya, and John Unsworth. 2013. "HathiTrust Research Center: Computational Access for Digital Humanities and Beyond". In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 395–396. JCDL '13. New York, NY, USA: ACM. https://doi.org/10.1145/2467696.2467767

Ponce, Mariano. 1912. *Sun Yat Sen: El Fundador de La República de China*. Manila: Imprenta de la Vanguardia y Taliba.

Prado-Fonts, Carles. 2018. "Writing China from the Rest of the West: Travels and Transculturation in 1920s Spain". *Journal of Spanish Cultural Studies*, April. https://doi.org/10.1080/14636204.2018.1453110

"READ | EADH – The European Association for Digital Humanities". n.d. Accessed 25 January 2019. https://eadh.org/projects/read

Saldaña, Zoë Wilkinson. 2018. "Sentiment Analysis for Exploratory Data Analysis". *Programming Historian*, January. https://programminghistorian.org/en/lessons/sentiment-analysis. https://doi.org/10.46430/phen0079

Ströbel, Phillip, and Simon Clematide. 2019. "Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images". In *Digital Humanities 2019*. Utrecht. https://doi.org/10.5167/uzh-177164

Tesseract-Ocr/Tesseract. (2014) 2020. C++. tesseract-ocr. https://github.com/tesseract-ocr/tesseract

"Texcavator". n.d. Accessed 25 January 2019. http://texcavator.hum.uu.nl/

"Text Correction Hall of Fame". n.d. Trove. Accessed 25 January 2019. https://trove.nla.gov.au/newspaper/hallOfFame?filter=newspaper

Tom. (2014) 2020. *Tmbdev/Ocropy*. Jupyter Notebook. https://github.com/tmbdev/ocropy

"Transatlantis Locations". n.d. Translantis. Accessed 25 January 2019. https://translantis.wp.hum.uu.nl/transatlantis-locations/

"Transkribus". n.d. Accessed 25 January 2019. https://transkribus.eu/Transkribus/

"Trove – Digitised Newspapers and More". n.d. Trove. Accessed 25 January 2019. //trove.nla.gov.au/newspaper

"Unsupervised MT: Fast and Accurate for More Languages". 2018. *Facebook Engineering* (blog). 31 August 2018. https://engineering.fb.com/ai-research/unsupervised-machine-translation-a-novel-approach-to-provide-fast-accurate-translations-for-more-languages/

Vanetik, Natalia, and Marina Litvak. 2019. *Multilingual Text Analysis: Challenges, Models, And Approaches*.

Viola, Lorella, and Jaap Verheul. 2019. "The Media Construction of Italian Identity: A Transatlantic, Digital Humanities Analysis of Italianità, Ethnicity, and Whiteness, 1867–1920". *Identity* 19 (4): 294–312. https://doi.org/10.1080/15283488.2019.1681271

"Welsh Newspapers Online – Home". n.d. Accessed 25 January 2019. https://newspapers.library.wales/

Wijfjes, Huub. 2017. "Digital Humanities and Media History. A Challenge for Historical Newspaper Research". *Tijdschrift Voor Mediageschiedenis* 20 (1): 4–24. https://doi.org/10.18146/tmg20277

Willems, Marieke, and Rossitza Atanassova. 2015. "Europeana Newspapers: Searching Digitized Historical Newspapers from 23 European Countries". *Insights* 28 (1): 51–56. https://doi.org/10.1629/uksg.218

"Xtas, the EXtensible Text Analysis Suite – Xtas 3.4 Documentation". n.d. Accessed 29 January 2019. http://xtas.net/

Zosa, Elaine, and Mark Granroth-Wilding. 2019. "Multilingual Dynamic Topic Model". Edited by Galia Angelova, Ruslan Mitkov, Ivelina Nikolova, and Irina Temnikova. *RANLP 2019 – Natural Language Processing a Deep Learning World*, International conference Recent advances in natural language processing, September, 1388–96. http://lml.bas.bg/ranlp2019/proceedings-ranlp-2019.pdf. https://doi.org/10.26615/978-954-452-056-4_159