The Forgotten Colonies: Building Corpora Containing Newspapers Published in German Colonies in German South West Africa and in Kiaochow from 1898 to 1914

Enhancing the Humanities

Not long after the unification of Germany in 1871, the German Empire started its colonization in Africa and in the Pacific, including German Southwest Africa (today's Namibia), German Kamerun (today's Cameroon), Kiaochow in Imperial and Early Republican China (today's Qingdao), etc. As part of its colonial practice, German government officials established newspapers that were designed for settler communities between the end of the nineteenth century to World War I. These publications offer immediate access for scholars to study colonial rule on military, political, economic, and cultural levels. The Forgotten Colonies is a Level 1 project and the main goal of this project is to create two customized corpora containing newspapers published in German colonies from 1898 to 1914. These two corpora allow future researchers to build projects using text analysis methods to examine the main themes presented in print products in German colonies both in Africa and in Asia.

This Level I project offers insights to a variety of scholars who work in the fields of digital humanities, German studies, African studies, Chinese studies, and media studies. On the technical level, the proposed project involves data collecting, data transferring (using OCR technique), data cleaning as well as testing topic modeling. All selected newspapers are currently available in pdf format stored in four archives in Germany, China, and Japan (the African collection is available online at the World Newspaper Archive through Michigan State University library license). Collecting, cleaning, and preparing data (transforming pdf files to text files) will be the central focus of this project. The scale of this project is relatively large, including six newspapers published in German Southwest Africa (Keetmanshooper Nachrichten, Keetmanshopper-Zeitung, Lüderitzbuchter Zeitung, Südwest, Swakopmunder Zeitung, and Windhuker Nachrichten) and four in Kiaochow, China (Amtsblatt für das Deutsche Kiautschou Gebiet, Deutsch Asiatische Warte, Tsingtauer Neueste Nachrichten, and Kiautschou Post). To ensure a fair comparison among publications between two regions, the time frame is set between July 1st, 1898 to December 31st, 1914. Digitization of newspapers in the African context is more advanced than in Asia. The World Newspaper Archive has all six publications stored and can be downloaded as pdf files. The total amount of available newspapers exceeds 2500 issues. The situation in Asia is less ideal. Many copies were destroyed and lost due to an unstable political situation from the late nineteenth century to the early twentieth century. The International Newspaper Museum in Aachen, Germany, Qingdao Museum in China and Kobe University in Japan all have partial copies. These copies are not available for download and will need to be requested. The materials in Asia are scattered and not well digitized. This project offers a complete, organized, and cleaned up collection of newspapers, which can be later used for text mining and other digital projects.

Multilingual, interdisciplinary, collaborative, and open-access are the guiding values of this project. The corpora will be in German and Chinese. All newspapers published in Africa

only used German while publications in Asia had both German and Chinese. The description of this project and the trail project using topic modeling will be written in English. This project builds on scholarly work conducted by African studies historians, German sinologists, Chinese historians who study newspapers, and text analysis specialists. Archival textual materials are collected from three institutions across the world. The outcome of this project will be a website (hosted on GitHub, presented through Git Pages), where scholars and interested general public have full access to the trail project analysis and both corpora (per request without university affiliation requirement). Keeping corpora and research results accessible to all parties will help bring scholar communities together through breaking institutional barriers. Being transparent and open on each step of this research helps scholars in the field to join the conversation and promote discussions on not only text analysis assisting research but also on colonial studies. Additionally, hosting this research on GitHub allows scholars, educators, students, and the general public around the world to have easy access to all materials. The website will also provide a channel for questions and feedback.

Environmental Scan

This project is developed based on scholarly works in three main fields, historian work in German South West Africa, sinological research in Imperial and Early Republican China, and Media studies (especially print publications). Through a photographic album produced by the German police in colonial Namibia, Lorena Rizzo's work studies police and prison institutions in the German colony (Rizzo). Pöppinghege wrote on the colonial press in Africa from 1898 to 1914, including the founding of the *Windhoeker Anzeiger* in German Southwest Africa, which is also included in the proposed project (Schäfer, 683). Historian Gertrud Pfister works closely with colonial newspapers and her work mainly revolves around physical activities and education (Pfister). All above mentioned works take a qualitative approach and focus on specific narrative telling.

Similar approaches can be seen on the Chinese end as well. Lin's master thesis provides an overview on newspapers published between 1897 to 1919 and does not go into details. Gao's work elucidates the German colonial policies in Kiaochow and the diplomatic negotiations between Imperial China and German Empire through close reading of three newspapers. From the media perspective, Zhou and Liu's work on German-Chinese bilingual publications illustrates the communicative, educational and community bonding functions that the print press served in Germany's short-lived colonial history in China (Zhou). Most scholarly works in this field are conducted with close reading techniques and provided limited examples from the archival materials. The only exception is Niu's book, *The Research on Der Ostasiatische Lloyd: 1886-1917*, where she takes a quantitative approach (using excel and SPSS) and analyzes the scale of the newspaper, most mentioned themes, advertisements, etc. Her usage of statistical tools are based on her initial close reading and categorization, which is still very different from the computational methods (topic modeling) proposed in this project.

Researchers have conducted projects using text analysis methods on German-langauge press within the field of digital humanities in the United States. Yet most scholarly works were

conducted based on existing corpora. Very limited research has been done from building corpora from the scratch. Jana Keck at German Historical Institute takes a data feminist approach to study social network of German-Americans. She challenges the algorithms that determine the relevance, which affects the result at the end. Her work aims at depicting the often neglected group: migrant women and discussing their role in the larger historical context. Keck's work is conducted based on an existing database, C19 German-American newspaper and the C21 digitized corpus of the newspapers (published between 1830 to 1914). This proposed project focuses on building two customized corpora from the very beginning, since both countries (Namibia and China) do not have digitized corpora for research use. Due to the unstable political situation between the late nineteenth century to 1949, local Chinese institutions do not have a complete collection of the required publications. Therefore, the success of this project requires collaboration between institutions in China, Germany, and Japan who each own a partial collection of the newspapers.

History of the Project

This project is inspired by Tianyi Kou's research paper "*Turnen* between Anti-Imperialist Resistance and Imperialist Expansion," written in March, 2021 under the guidance of Prof. Matthew Handelman. Kou's research interest in sports' role in German Southwest Africa generated from her conversations with Prof. Peter Alegi in 2020. In her previous research, she selected three newspapers and primarily focused on illuminating the role of physical activity, *Turnen*, and its relationship to nationalism in the colonial context. In 2021, Kou decided to look into the print press in another colonie in Asia and to expand her research into a comparative study between two regions. Being a native speaker of Chinese, trained in English and German, she acquires the ability of reading scholarly work in all three languages, which allows her to gain a more complete view on her research subject. All research work between 2020-2021 is conducted under the support of units at Michigan State University (MSU), such as the College of Arts and Letters, and the MSU Library. The computational aspect of this project is supported by the Computational Humanities (CH) Team at Leipzig University in Germany. The research will be hosted on GitHub as well as on the MSU Domain. Although this project is not directly related to Kou's doctoral dissertation, she plans to carry it with her to her postdoctoral facility.

Activities and Project Team

This project is planned to be completed in 24 months. In Phase I (month 1-18), the tasks for the research team are to 1) contact and acquire newspapers from three institutes in Germany, China, and Japan 2) collect and categorize newspapers published by ten presses in two regions 3) use ABBYY to transfer newspapers published in Africa (all in German) from pdf files to machine-readable text files 4) use ABBYY to transfer newspapers published in Asia (most in German, some in Chinese) from pdf files to machine-readable text files. In Phase II (month 19-24), the research team will 1) run a pilot project with smaller corpus (50 issues), test the model, adjust the code based on pilot project's result 2) upload both corpora to this project's website 3) write a research paper on the process of building multilingual corpora 4) peer review

and submit for publication. The research team also plans to document the entire process from collecting and cleaning data to running a pilot project until writing the analysis.

Project staff are Tianyi Kou, Computational Humanities (CH) team members at Leipzig University, five to seven undergraduate students at MSU, and a freelance web designer. Kou and undergraduate students will be the main blog writers in Phase I; CH team members will take over the blog post duty once graduate students complete their duties after month 18. Five to seven undergraduate students are hired to assist data preparation and cleaning in Phase I. Through project-based learning and blog writing, students will gain not only experience in learning and utilizing corpora building techniques, but also in research ethic and collaboration skills. Kou and CH team members will benefit from mentoring undergraduate students as well as cross-institutional collaboration. At the end of this project, all team members will discuss the entire research process and provide constructive feedback. The research paper will be open for peer review starting from month 22. Potential peer reviewers are digital humanists working in text analysis (specifically in corpus building) and colonial scholars in both African and Chinese contexts (such as Gertrud Pfister, Klaus Mühlhahn, Yi Zhou, etc.).

Final Products and Dissemination

As stated previously, the final products include two corpora (in German and Chinese), a detailed analysis written in English, and a research journal in blog style written in English. All above mentioned items will be hosted on two identical websites, one through GitHub and the other one on Michigan State University College of Arts and Letters domain. The two corpora are available to interested scholars who plan to build research on this project. Educators in colonial studies (both in African and Asian context) as well as in digital humanities are encouraged to integrate the research results and methods into their own teaching practice. The general public also has free access to research materials. Furthermore, the research team will ensure that all textual materials are suitable for text to speech softwares.

After the project is completed in month 24, Kou will submit the written analysis to journals for publication. Once accepted, all team members (team leader, CH group members, undergraduate students) except for the web designer will be listed as co-writers. The final project will also be submitted for conference presentation at the German Studies Association Annual Conference, Annual Association for Interdisciplinary Studies Conference, African Studies Association Conference, etc. The web designer will be paid for their service and their name will also be mentioned on the websites.

Work Plan

Team Members:

Tianyi Kou (key person), Computational Humanities Team members (2 people), undergraduate students (5-7 people), a freelance web designer

Work Plan in Details

	Time Frame	People Involved	Activities and Evidence of Outcomes
Phase I Corpora Building	Month 1 - 18	All team members	Two completed corpora: Corpus A should contain six newspapers (2749 issues) published in German South West Africa Corpus B should contain four newspapers (app.800-1000 issues) published in Kiaochow (Corpora are available per request)
Step 1: Data Collection	Month 1 - 3	Key person, undergraduate students, web designer	Key person works with the web designer on setting up the website. Key person reaches out to three institutions with archival materials. Five undergraduate students organize and categorize required newspapers. A complete collection of all 10 newspapers (in pdf format), stored both locally at two places and in the cloud.
Step 2: Data Conversion I	Month 4-10	CH team members, undergraduate students	Under the guidance of CH team members, undergraduate students (who can read German) use ABBYY to convert pdf files to text and store text in the first corpus. Key person and CH team members randomly select results and make sure the quality of the text meets the standard.
Step 3: Data Conversion II	Month 11- 16	CH team members, undergraduate students	Under the guidance of CH team members, undergraduate students (who can read German and Chinese) use ABBYY to convert pdf files to text and store text in the second corpus.

			Key person and CH team members randomly select results and make sure the quality of the text meets the standard.
Step 4: Data Preparation	Month 17-18	Key person, CH team members, undergraduate students	Key person and CH team members clean the first batch of data (50 issues) while undergraduate students observe and learn from them. Undergraduate students then apply their skills in practice and finish all data preparation.

	Time Frame	People Involved	Evidence of Outcomes
Phase II Comparison, Analysis, and Peer Review	Month 19 - 24	All team members	A paper documenting the corpora building process (in publication-ready quality); A research journal in blog style written in English (Both will be hosted on a website.)
Step 1: Pilot Project & Reflection	Month 19 - 20	Key person, CH team members	Key person and CH team members conduct a pilot project where they take 50 issues of newspaper from Corpus A and run a topic modeling program. Based on the result, researchers will adjust the code.
Step 2 Conclusion Writing	Month 21-22	Key person	Key person writes a paper on the corpora building process and publishes it on the website.
Step 3 Peer Review & Final Wrap-up	Month 23-24	Key person, web designer	Key person works with the web designer and publishes research materials on the website. Key person sends out invitations for peer review. Peer review process takes 4-6 weeks. Key person then revises the written materials before submitting for publication.

Budget Estimation

Estimated Expenses	People Involved	Amount (in U.S. dollars)
Key Person annual compensation	Tianyi Kou (24 months)	7500
Computational Humanities team members stipends	Computational Humanities team members (2 people, 18 months)	4500 x 2 = 9000
Undergraduate students wages	Undergraduate students (5-7 people, 18 months)	3500 x 7 =24500
ABBYY software purchase fee	All team members	199 (windows) x 5= 1194 129 (mac) x 5 = 645
Web-designer service fee	Web-designer (4 months)	3000
Cloud-based storage fee	None	400
Hard drives for storage use	None	400
Publication costs	Team	300
Miscellaneous	Team	100
	Total Expense	**47039**

Data Management Plan

Roles and Responsibilities

The primary sources of this project are ten newspapers published between July 1st, 1898 to December 31st, 1914 in German Southwest Africa and in Kiaochow. The six newspapers published in Africa are acquired through the World Newspaper Archive under Michigan State University license. The four newspapers are currently partially hosted by Qingdao Museum in China, the International Newspaper Museum in Germany, and Kobe University in Japan. With the permission of the above-mentioned institutions, all team members have the right to read and study materials, but are not allowed to modify the material content or make copies outside of the research project use. All materials will be maintained in digital forms and will not be hosted on the project website. The two final corpora that contain textual data are available to the public per request.

Expected Data

Data involved in this project include 1) primary sources collected from archives 2) two corpora, 3) the result generated from corpus I 4) code used in the text analysis process and 5) a written paper documenting the corpora building process. Depending on archives' requirements, item 1 will only be accessible to team members during the 24-month research period. Item 2 and 4 are available to scholars and educators upon request. Item 3 and 5 will be published on the project website.

Items 1, 2, and 4 will be stored in two devices locally as well as through a paid Cloud service on the Cloud. Items 3 and 5 will be presented through the project websites, one hosted on GitHub and the other one through Michigan State University College of Arts and Letters domain.

Period of Data Retention

The project director will maintain the project data every 45 days and ensure data availability to scholars, educators, and the general public. The project director will also carry the project with her in the case of leaving her current institute (remove materials from MSU domain).

Data Formats and Dissemination

Item 1 will be in pdf format. Item 2 will be in two .txt files. Item 3 and 5 are in markdown format. All team members except for the web designer have access to all data. Item 2 is available for users upon request. Users who request these two corpora will be documented.

Data Storage and Preservation of Access

The team will purchase two external hard drives for storing the data. All data will also be stored via a paid Cloud-based service. Item 3 and 5 will be hosted both through GitHub and MSU domains.