Multilingual Computational Analysis of Historical Texts:
A Case Study of *Robinson Crusoe* Translations

Annice Chang, Kaeshav Krishna, Yuna Miyoshi, Amy Shi, Ziqiao Wang & Dr. Ali Bolcakan, Prof. Christi Merrill

Introduction

In the past two decades, there have been great strides in the realms of text processing and analysis, machine learning and translation, and large language models. Still, working with historical and multilingual textual data remains a major challenge because of linguistic variations and related shortcomings in text recognition and parsing. Natural Language Processing (NLP) relies on high-accuracy Optical Character Recognition (OCR), which is particularly difficult with historical and multilingual datasets. Thus, we seek to do our experiments without relying on an understanding of the textual material but instead intend to identify reusable semi-stable anchor points for our comparative computational analyses.

Taking Daniel Defoe's English language novel *Robinson Crusoe* (1719) as our test case because of its significant impact and availability in translation, we look at a variety of Japanese, Mandarin, and Tamil translations from the late 19th and early 20th centuries, sourced through humanistic archival research. Our primary objective is to come up with reproducible computational frameworks, methods, and tools to analyze the distinctions between the source text and its translations, even with non-Latin script material in under-resourced languages.

This foundational work should facilitate the discovery of additional resources and enable analyses of variations in tone, style, and narrative shifts across languages in historical textual data. This project seeks to contribute to both traditional and computational humanities, translation and reception studies, computational linguistics, and computer and information sciences.

Workflow & Methods

Data acquisition:

After we settled on *Robinson Crusoe* for our experiment, we first started doing humanistic, archival research. We began collecting Chinese, Japanese, and Tamil scholarship on the histories of translation of *Robinson Crusoe* in these languages and cultures and broader research on *Robinson Crusoe.* Therefore, we knew that the majority of the translations of *Robinson Crusoe* would be after 1900, almost 3 centuries after its first publication (see figure below). Based on our findings, we compiled lists of translations to search for. Many of these works were simply unavailable (or undiscoverable because of metadata issues) in the largest digital repositories in the English-speaking world: Google Books, Internet Archive, and HathiTrust, and smaller databases accessible through the U-M library. We painstakingly collected translations from international databases as much as we can but realized that were still missing scores of work we knew to exist. This process showed us how difficult it is to conduct multilingual digital humanities, as digital archives are usually not connected to each other with linked data, and most platforms have their non-standardized way of generating and handling metadata.

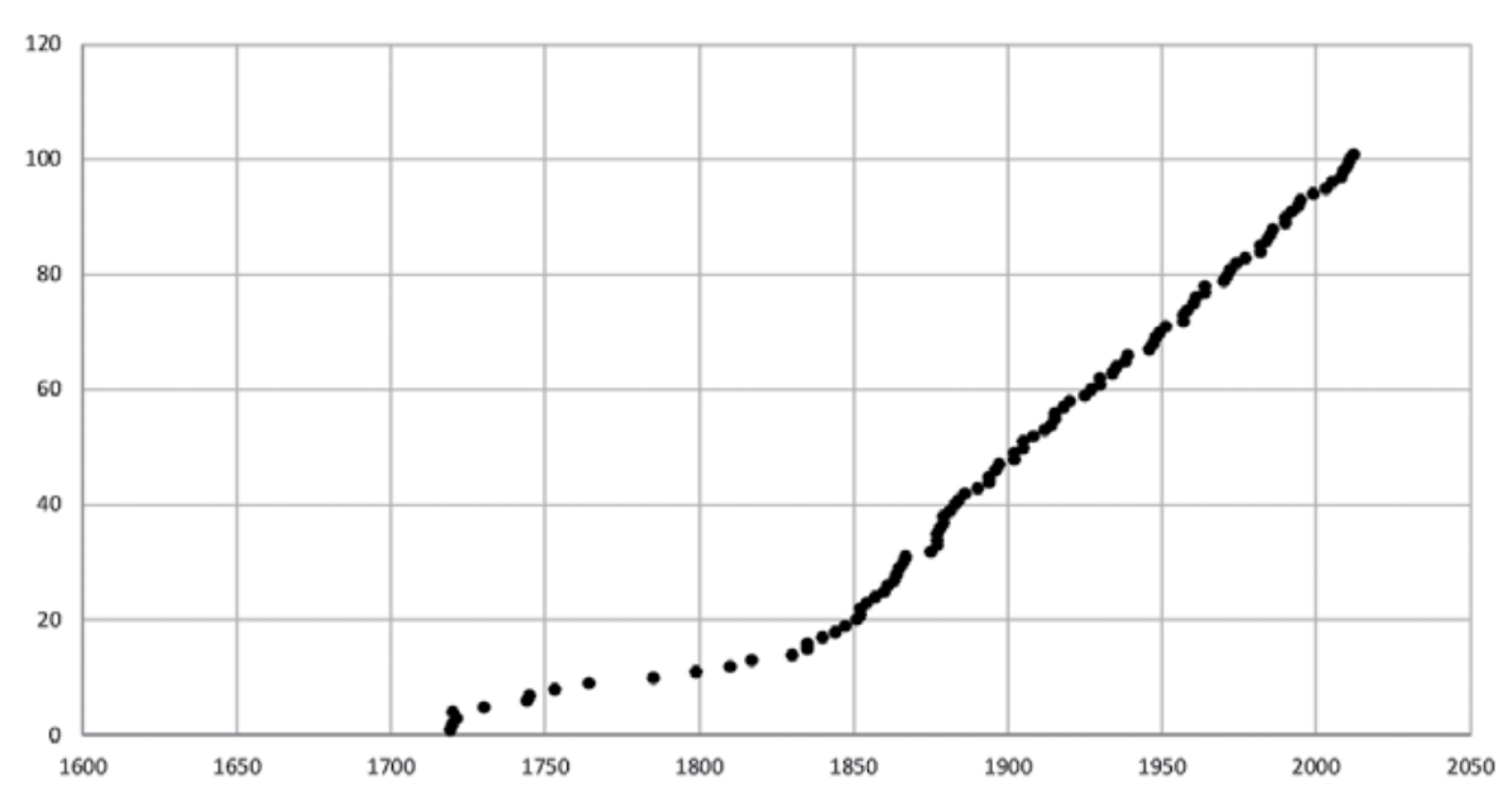
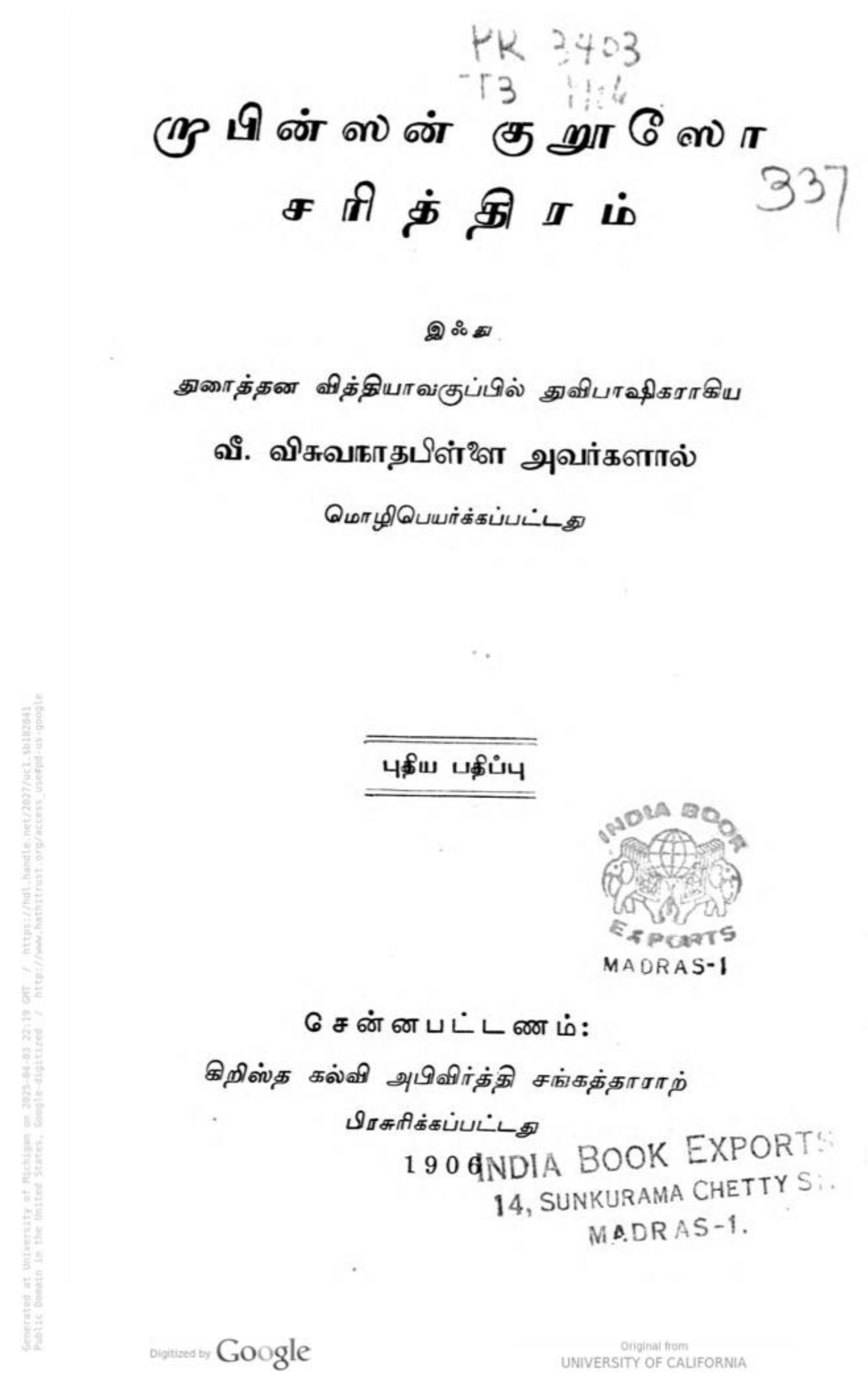


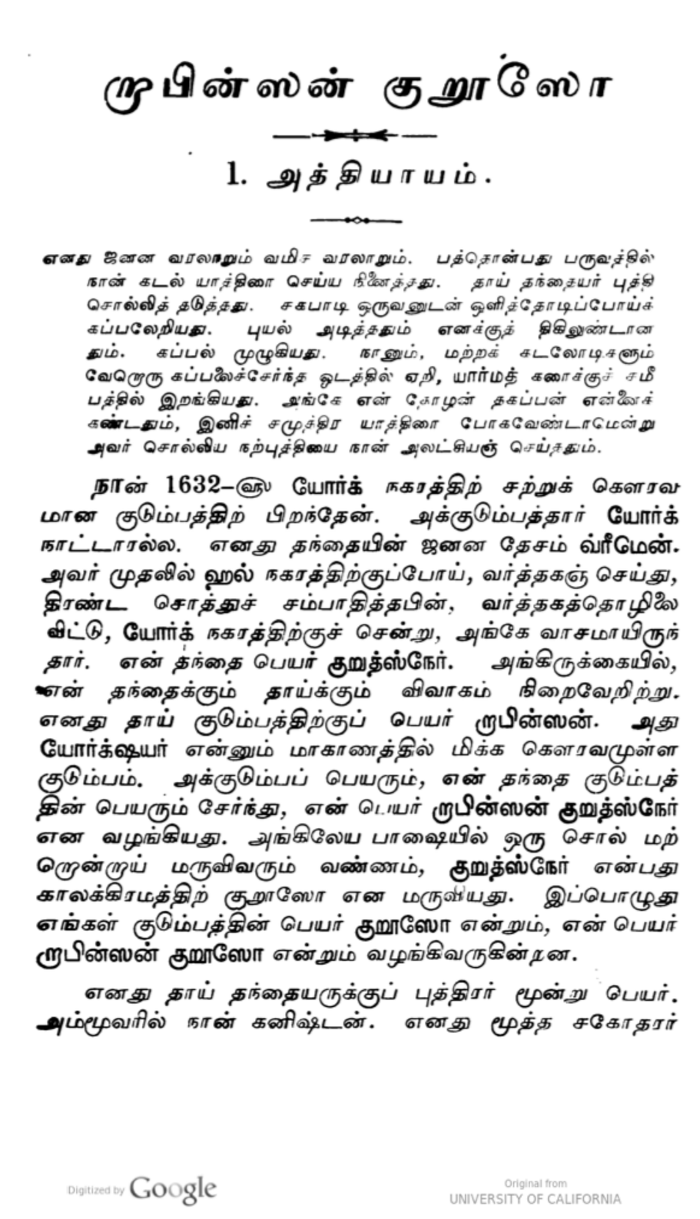
Chart 3 First print translations of *Robinson Crusoe*

of

Cited from Peter Hill, "Translation and the Globalisation of the Novel: Relevance and Limits of a Diffusionist Model." *Migrating Texts: Circulating Translations around the Ottoman Mediterranean*, edited by Marilyn Booth, 2019, pp. 95-121.

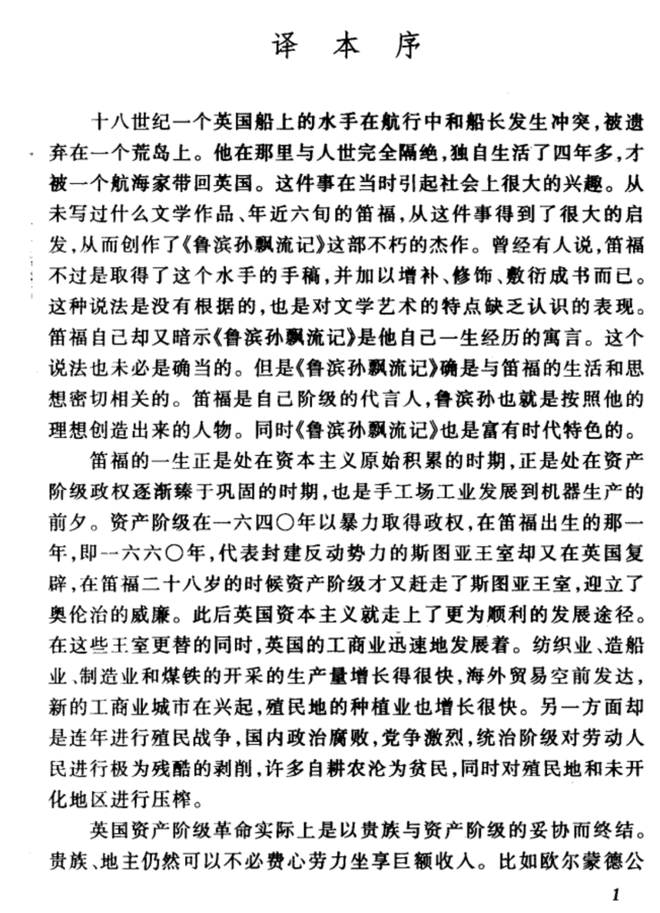
Data Processing:





Tamil is a Dravidian language with its own unique abugida script with 247 characters. We found two Tamil translations, from 1906 and 1915, on HathiTrust. Thus they had a workable OCR layer.

The challenge we couldn't overcome was that while Tamil is an official language in India, Singapore, and Sri Lanka, we couldn't locate a rich digital repository with other translations needed for our comparative analysis.



While HathiTrust had a few Chinese (Mandarin) sources, we found many other Chinese translations from international Chinese language repositories. While OCR for post-1949 Simplified Chinese is fairly robust, the same can't be said for pre-1949 texts in Simplified chinese. Script variations, annotations, and the condition of these older books, coupled with low-scan quality for some books, presented important challenges.

Working with early versions of Japanese texts presented multiple challenges . Unlike modern Japanese, Meiji-era (1868–1912) included use of obsolete kanji and orthography that most OCR engines failed to recognize, and many texts featured handwritten, "slurred" characters, similar to cursive, where strokes blended together. These combined factors resulted in creating significant OCR errors where characters were misread, omitted, and added., rendering automated text extraction unreliable for our research.

Script

Using Python, we parsed through each line in the file, using the ".split" and ".strip" functions to create a mini-list for each line, called "words." We then added each instance of "words" to a main list called "words-per-line." The "words-per-line" list was eventually used to populate a pandas data frame. This data frame created a matrix-like structure through which we could easily track the location of a word in the text, down to which line and which word in the line the desired word was located. From there, it was a simple matter of parsing through the data frame and printing the exact location of the desired word as a tuple. The same process, reproduced in the English version, gives us exact tuples for word locations, setting the stage up for a series of semi-stable anchor points to eventually lead us to our desired result.

```
                     0            1            2  ...    11    12    13
0                 This         file          was  ...  None  None  None
1                 Find         more        books  ...  None  None  None
2                 None         None         None  ...  None  None  None
3               Title:    Rāpinsan       Kurūsō  ...  None  None  None
4   molipeyarkkappaṭṭatu         None         None  ...  None  None  None
..                 ...          ...          ...  ...   ...   ...   ...
68     காலக்கிரமத்திற்     குறூஸோ        என ... None   None   None
69        எங்கள் குடும்பத்தின்       பெயர் ... None   None   None
70        றாபின்ஸன்     குறூாஸோ      என்றும் ... None   None   None
71          எனது      தாய் தந்தையருக்குப் ... None   None   None
72        அம்மூவரில்       நான்     கனிஷ்டன். ... None   None   None

[73 rows x 14 columns]
Word found at: [(54, 1)]
               0          1          2            3  ...    15    16    17    18
0            The    Project  Gutenberg        eBook  ...  None  None  None  None
1           None       None       None         None  ...  None  None  None  None
2           This      ebook         is          for  ...  None  None  None  None
3           most      other      parts           of  ...  None  None  None  None
4    whatsoever.        You        may         copy  ...  None  None  None  None
...          ...        ...        ...          ...  ...   ...   ...   ...   ...
10540       None       None       None         None  ...  None  None  None  None
10541       This    website   includes  information  ...  None  None  None  None
10542  including        how         to         make  ...  None  None  None  None
10543    Archive Foundation,       how           to  ...  None  None  None  None
10544  subscribe         to        our        email  ...  None  None  None  None

[10545 rows x 19 columns]
Word found at: [(72, 6)]
```

Analysis

The original *Robinson Crusoe* text had approximately 123936 words. We estimate that the Chinese translations would have an average multiplier of 1.5, 2.5 for Japanese , and 1.2 for Tamil. We planned to use these as benchmarks for our comparative analyses and produce new ones that would be specific to historical and literary variations of these languages. While we didn't get to finish our experiment, we did confirm our initial hypothesis. We can still posit that historical translations tend to be markedly, sometimes radically, different than the source text. Because the European/American novel as a cultural technology was sort of a new device in other contexts, translators had to make varying levels of changes to conform to the demands of the book market, social morals, and political realities.

Future Directions

For the next stages of the project, we strongly believe that it'll be necessary to pivot to machine-learning-based Vision OCR. We might have to adjust the models to increase the accuracy for literary and early 20th century texts, both of which are still barriers in the Vision OCR domain. We will also have to switch to clusters of anchor-points for more accurate matching and avoiding false positives.

Acknowledgements

We are grateful to the wonderful librarians and research staff of the U-M Library: Kat Hagedorn, Director, Digital Content & Collections; Yunah Sung, Korean Studies Librarian; Liangyu Fu Director, Asia Library; Keiko Yokota-Carter, Japanese Studies Librarian; Lara Deanna Unger, Digital Conversion Supervisor.