An attempt to overcome language barriers in library collections through AI:

University of Hawai'i at Mānoa Library's Turkish Brigade Archival Collection

Ellie Kim (Korean Studies Librarian), Boyoung Choi (Korea Foundation Intern), Yeajin Park (Visiting Librarian)

Introduction

Korean War

Korean War, conflict between the Democratic People's Republic of Korea (North Korea) and the Republic of Korea (South Korea) in which at least 2.5 million persons lost their lives. The war reached international proportions in June 1950 when North Korea, supplied and advised by the Soviet Union, invaded the South. The United Nations, with the United States as the principal participant, joined the war on the side of the South Koreans, and the People's Republic of China came to North Korea's aid. After more than a million combat casualties had been suffered on both sides, the fighting ended in July 1953 with Korea still divided into two hostile states. Negotiations in 1954 produced no further agreement, and the front line has been accepted ever since as the de facto boundary between North and South Korea.

Türkiye(Turkey)'s participation in the Korean War

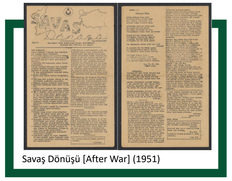
At the United Nations Security Council held at the request of the United States, North Korea's military actions were defined as aggression, and the United Nations therefore requested North Korea to stop its military actions and withdraw from South Korea. As North Korea refused, the UN decided to send troops to the Korean War. In accordance with this decision, the Turkish army entered the war as part of the UN forces and was one of the four countries that sent a large brigade-level army. Turkey was the second country to decide to enter the war, following the United States. During the Korean period, Turkey dispatched 21,212 people and in the process of carrying out various battles and operations, 966 people were killed, 1,155 people were injured, and 244 people were taken prisoner. This was the fourth largest number of casualties among the 16 countries of the UN forces that participated in the war.

University of Hawaii at Manoa Library's Turkish Brigade Archive

The Korean Collection at UHM Library has recently acquired the personal archive of Major Şükrü Doğan, a Turkish military officer who served in the Korean War. The archive includes the following materials.

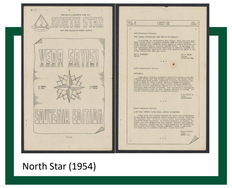
Korea ve Harp [Korea and War] (1954)

Korea ve Harp is a mimeographed publication from Turkish Brigade. It is divided into various sections, each contains with information and statistics such as discourse on Korean history, religion, geography, an explanation of the Korean alphabet, history of the Korean War including details on the protagonists and specific battles, maps of Korea, lists of the Turkish Brigade soldiers then serving in theatre, Turkish fallen soldiers, and the locations of the fallen soldier's graves in the cemetery in Pusan.



Savaş Dönüşü [After War] (1951)

Savaş Dönüşü is a broadside magazine mimeographed by members of the Turkish Brigade aboard the ship, USS General C.C. Ballou, traveling en route from Korea to Turkey. Each issue features news stories from home and reports on the war action in Korea acquired from the ship's radio, as well as short articles, poems, humorous vignettes and inspiring quotations. It follows a tradition of soldiers publishing their own periodicals during long sea voyages, which became popular during the World War I era. This archive includes four issues of Savaş Dönüşü.



North Star (1954)

This magazine was mimeographed by members of the Turkish Brigade aboard the ship USS General R. M. Blatchford, en route from South Korea to Turkey. The magazine is named 'North Star' (Turkish: Şimal Yıldızı or Kutup Yıldızı) after the code name of the Turkish Brigade. In line with similar soldiers' periodicals published aboard ship during long voyages, North Star features amusing and morale-boosting short stories, poems, patriotic quotations and humorous vignettes, as well as reports concerning the journey itself.

Birinci kafile ile yurda gidecek Asb. Isim cetveli [Soldiers who will Return Home with the First Group / Name Table] (1951-1954)

- A mimeographed list of Turkish Brigade soldiers who were scheduled to imminently return to Turkey. The printed list features 151 soldiers and 1 name added at the end in the manuscript. The name of each soldier is accompanied by their serial number, unit, rank, and tag number.

Yurtta sulh, cihanda sulh." [Peace at Home, Peace in the World]

- An original photomontage made for an anonymous Turkish Brigade soldier, depicting his portrait, a map of the Korean War theater, and the figure of a lady, adorned with a Turkish flag and carrying a torch, personifying 'peace', following a common motif in both contemporary Turkish and international iconography.

Original Turkish Brigade Photographs (1953-1954)

- Collection of original photographs taken by members of the Turkish Brigade, mainly in Korea (variously in Seoul, Pusan, and along the battle front), but also while 'on leave' in Tokyo, Japan, as well as during the homeward sea voyage between Korea and Turkey.
- These are primary sources produced during the Korean War and are rare, especially those written by soldiers from non-English speaking countries. They are expected to shed light on the situation of the Turkish army at that time, which will be of great help in the study of the history of the Korean War.

Challenge

The challenge is that all items of the collection are written in Turkish, even though the prime users are Korean and English speakers

Solution

We decided to create a process using AI.
We also expected the effects of time saved and cost reduction.

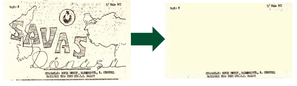
OCR(Optical character recognition)

Subject Data

115 items of handwriting annotations on the verso of the photographs, 4 items of mimeographs (unrecorded magazines, name list). Created digital image files of them by scanning

Preprocessing

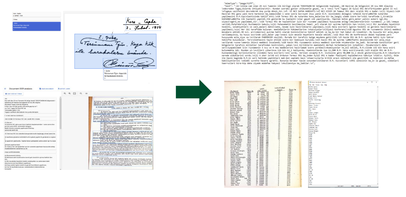
Separated vertical columns and blocks,



Erased illustrations of the magazines, Level the angle, and Adjust the brightness.
4 mimeographs divided into 98 items.

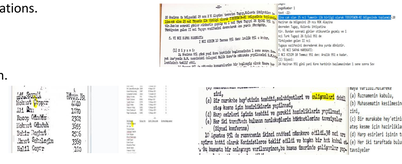
AI-OCR : Google Cloud Document AI

This is a document processing AI that features an OCR Processor and Handwriting Recognition in Turkish. It provides free credits that covers us to run the program. In return, we transformed this unstructured information into text data in JSON.



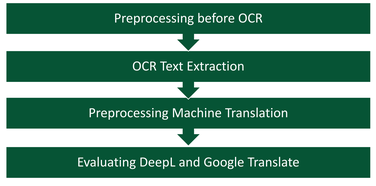
However, when we reviewed the output to move to the next step (translation), we found the following limitations.

- Line breaks and tables require an post-refinement procedure.
- Poor print quality led to poor recognition.
- Handwriting recognition was also poor.
- Failed to recognize diacritical marks.
- Time consuming post-processing

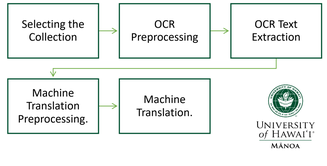


Reviewing the OCR output corresponds to the preprocessing work for the translation. We refined the outcome within the abovementioned limitations and generated source text files for machine translation.

Process



Processing Model



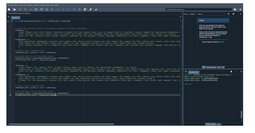
Machine Translation

The JSON file obtained from OCR text extraction went through a preprocessing process for machine translation. While using OCR, there were a few cases where line breaks happened in the middle of a sentence. In these cases, machine translation was not natural, so we took caution and went through the preprocessing process.

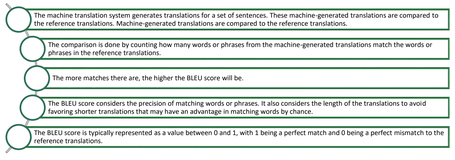
Human Translation	DeepL	Google Translate
Even before the first written historical accounts of the region, there was a people with a neolithic culture, formed on agricultural foundations in Korea. The ancestors of this people were a mixture of tribes, who had migrated from the Siberian Plateau, and rivers and river basins in China. Due to the isolation from outside influence afforded by the peninsula, Koreans developed differently to their neighbours, China, and Japan, developing the distinct language and physical appearance, which we see today.	On agricultural foundations in Korea, even before the first written history of the regionThe people of this nation were mixed tribes who migrated here from the Siberian steppes and the basin of the Chinese rivers, and thanks to the Peninsula's ability to protect this nation from the outside world, the Koreans developed in a different way from their neighbors, the Chinese and the Japanese, and acquired the distinct language and physical appearance they have today	Even before the region's first written history, there was a history of agricultural foundations in Korea. There was a nation with a neolithic (Polished Stone Age) culture. The ancestors of this nation were mixed tribes who migrated here from the Siberian steppes and the Chinese river basin. Thanks to the peninsula's isolation of this nation from outside, the Koreans are different from their neighbors, the Chinese and Japanese. They have developed in some way and have acquired today's distinct language and physical appearance.

Because the original text is written in Turkish, it was impossible to confirm whether it had been translated accurately.
So, we decided to use BLEU score which can evaluate machine translation.

BLEU stands for "Bilingual Evaluation Understudy". It is a metric commonly used in natural language processing (NLP) for evaluating: In paraphrasing tasks, BLEU can measure the similarity between a generated paraphrase and the original sentence.



DeepL BLEU Score: 0.12445399860315087
GoogleTrans BLEU Score: 0.0974118872676338



Waheed, Ahmed (2023,May). How to calculate the BLEU Score in Python. Retrieved on February, 14, 2024 from https://theprithoncode.com/article/bleu-score-in-pythonHhow-to-calculate-the-bleu-score

In addition, we did a survey of native English speakers. To four evaluation criteria: whether it was fully translated into English, natural, appropriate words, and grammatical errors. Then we asked "Why did you choose this option?"

Here are the following answers:

Even though both paragraphs are grammatically correct and though the second paragraph seems shorter and more smooth, the first and last sentence seem unnatural and not what an L1 speaker would write. Moreover, the paragraph feels like it has more comma splices and uses "and" a lot which I don't think a translator would do.
1. The first sentence is incomplete. The sentence feels slightly long; there should probably be a period separating it somewhere 2. The last sentence "They have developed in some way" is a little awkward
I chose my answer mainly due to the run on sentence in the first paragraph, which I believe could have been broken up by placing a period in between "....Chinese rivers" and "...,and thanks". I think they were both translated well, but I preferred the second one based on this reason and sentence flow.
It is hard to mark this section ... I do not know if all sentences have been translated. There are not major grammatical errors...more like a case of better words could have been used and sentences could be shorter.

Conclusion

limitation	significance
Since the original text was in Turkish, which we cannot speak, it was impossible to check whether the machine translation was translated properly — we hired a professional translator.	AI OCR enabled us to create the ST(Source Text, a text written in a given source language which is to be) and fostered effective communication with professional translators, out of the original fragile material and unstructured data.
Reviewing the OCR results is laborious and consumes a significant amount of time.	Our endeavor can be considered as an illustration of surpassing the barrier of language.
We have confirmed that depending solely on AI for processing collections in unfamiliar languages is not flawless and necessitates laborious post-processing work.	We established a process for collecting third-party language resources that can be used in the future.