The Timbuktu Chronicles: Summary of a Digital Humanities Project

Ayah Aboelela

September 2021

Intro

The Timbuktu Chronicles are a series of manuscripts written in seventeenth century Timbuktu, generally discussing the history of West Africa. Two of the longest and most famous of these chronicles are *Tarikh al-Fattash* by Mahmud Kati and *Tarikh as-Soudan* by ʻAbd al-Raḥmān ibn ʻAbd Allāh Saʻdī. This digital humanities research project aims to explore these two chronicles via various software, making the texts, or selected passages of them, more accessible to audiences without necessarily high Arabic fluency. So far, the software tools used include named entity recognition (NER), optical character recognition (OCR), and diacritization. This paper will overview the various DH tools and libraries used on the Timbuktu Chronicles, as well as some collected results.

Translations

Currently, the two manuscripts are available to access, use, and modify in both the original Arabic and the French translation. However, since all English translations are under copyright, we tested some available machine translation tools. Using Google Translate, we translated the original Arabic into English and compared it with the translation from French into English, which appeared to produce more coherent results. However, DeepL's machine translation from French into English appeared even clearer, and so, with a few edits for comprehensibility and clarity, this is the tool we used to obtain all English translations of these chronicles.

Named Entity Recognition

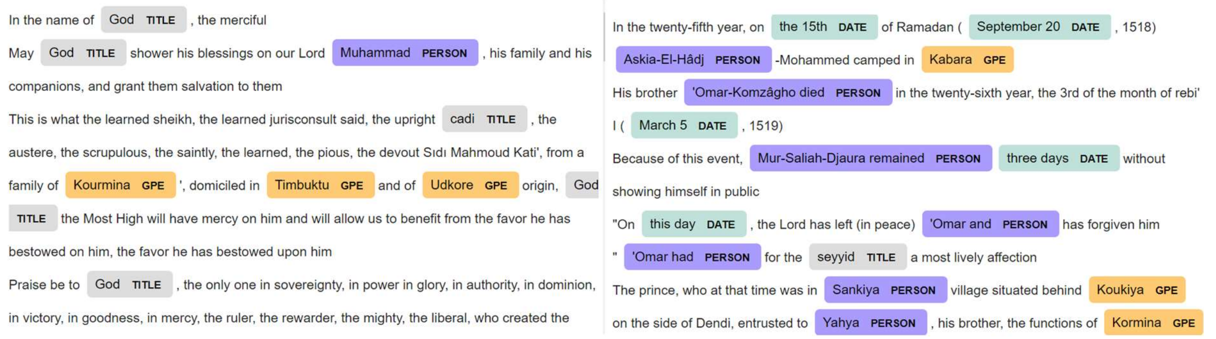
NER can be useful for identifying proper nouns and potentially important key figures, locations, tribes, and inter-textual references. For this project, I used SpaCy's library to train NER models which could label people's names, organizations, geopolitical entities, nationalities or religious or political groups (for example: Songhay people, Sunnis, Shias, etc.), and time indicators. The data consisted of 250 lines collected from both *Tarikh al-Fattash* and *Tarikh as-Soudan,* split into 200 lines for training, and 50 for testing.

Spacy offers a variety of models for NER, and I used three of them to determine which would be most suitable for the data. A table of each Spacy model along with an accuracy score for its NER labels are given below:

Spacy Model	Pre-Training Accuracy Score	Post-Training Accuracy Score
en_core_web_sm	0.59	0.75
en_core_web_md	0.60	0.82
en_core_web_lg	0.54	0.76

Though the difference was not significantly better for any one model, I decided to use en_core_web_md for NER Training. Going through the output, I realized that one cause for mis-labeling was that certain titles, in both Arabic and native West African languages, were being mis-labeled as people's names, geographic locations, or geopolitical entities. Some examples include shaykh, shaykh-ul-islam, cadi, koi, faran, and khalife among others. Therefore, I tried to add an additional label for the NER model: “title.” After adding this to the model and training it, the resulting accuracy on the labelled test data reached 0.83.

The following is a sample of an NER visualization, using SpaCy’s visualizer Displacy, on passages from the chronicles:



In the name of God TITLE , the merciful
May God TITLE shower his blessings on our Lord Muhammad PERSON , his family and his
companions, and grant them salvation to them
This is what the learned sheikh, the learned jurisconsult said, the upright cadi TITLE , the
austere, the scrupulous, the saintly, the learned, the pious, the devout Sıdı Mahmoud Kati', from a
family of Kourmina GPE ', domiciled in Timbuktu GPE and of Udkore GPE origin, God
TITLE the Most High will have mercy on him and will allow us to benefit from the favor he has
bestowed on him, the favor he has bestowed upon him
Praise be to God TITLE , the only one in sovereignty, in power in glory, in authority, in dominion,
in victory, in goodness, in mercy, the ruler, the rewarder, the mighty, the liberal, who created the

In the twenty-fifth year, on the 15th DATE of Ramadan (September 20 DATE , 1518)
Askia-El-Hâdj PERSON -Mohammed camped in Kabara GPE
His brother 'Omar-Komzâgho died PERSON in the twenty-sixth year, the 3rd of the month of rebi'
I (March 5 DATE , 1519)
Because of this event, Mur-Saliah-Djaura remained PERSON three days DATE without
showing himself in public
"On this day DATE , the Lord has left (in peace) 'Omar and PERSON has forgiven him
" 'Omar had PERSON for the seyyid TITLE a most lively affection
The prince, who at that time was in Sankiya PERSON village situated behind Koukiya GPE
on the side of Dendi, entrusted to Yahya PERSON , his brother, the functions of Kormina GPE

Optical Character Recognition

To work with the chronicles, we need pure text files containing the French and/or Arabic text. As the chronicles currently exist on HathiTrust as PDF files, we used HathiTrust’s built-in OCR tool (which uses Google’s OCR engine) to get the text files. For the French version, this was mostly non-problematic, as the resulting text was readable and accurately reflected the PDF images. However, the OCR Arabic text obtained from HathiTrust contained several errors. One reason may be that *Tarikh al-Fattash* used an older font that changed the number of dots belonging to certain letters, as well as positioning some dots from above the letter to below the letter, both uncommon conventions today. This may be reflected in the error rates, which tend to be higher for *Tarikh al-Fattash* than they are for *Tarikh as-Soudan.*

To compare various OCR tools, I used Simorgh’s OCR platform to produce Arabic text based on images of the Timbuktu Chronicles’ PDF files. I tested the output of HathiTrust’s OCR against Simorgh’s output, comparing each of them with the manually corrected Arabic text. I then measured their error rates on both the character level and the word level, using the Fastwer library. The results are as follows.

Word Error Rate (WER):

Text / OCR Tool	HathiTrust	Simorgh
Tarikh al-Fattash	27.5	26.4
Tarikh as-Soudan	18.3	22.3

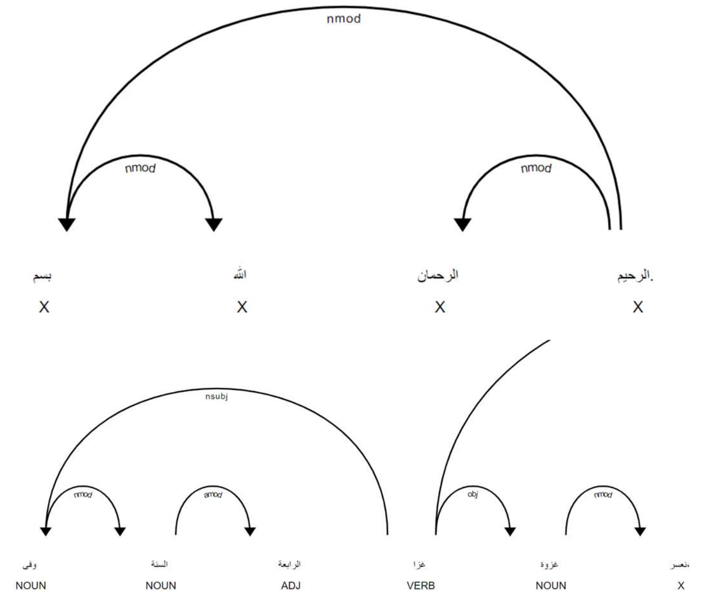
Character Error Rate (CER):

Text / OCR Tool	HathiTrust	Simorgh
Tarikh al-Fattash	6.3	5.6
Tarikh as-Soudan	5.1	5.6

The scope of this project did not include training OCR models, however, in the future, it would be useful to do so for the particular fonts used in these Timbuktu Chronicles to better improve the resulting text produced from the images.

Dependency Parsing

Another tool useful for understanding sentences in Arabic is dependency parsing. For this, I used Stanza, an NLP package which also supports Arabic models. I was also able to integrate Spacy's visualizer, Displacy, to display Stanza's dependency parsing output in an easily traceable format. The drawback is that the words are listed from left to right, instead of right to left as it is in the original Arabic. A sample for this is shown below.



Disambiguation

The original Arabic texts do not include diacritical marks for most words, since it is generally expected that readers of such texts would be adequately familiar to understand what the diacritics are based on context. However, for new or unfamiliar Arabic speakers/readers, it is useful to have the diacritics available for ease of reading, learning, and understanding. CAMeL Tools, a group of tools for Arabic NLP, provides tools for disambiguation, the process of adding diacritics to unmarked Arabic words. This tool is generally accurate and, despite a few errors, makes the text much more readable with its predicted diacritics. A sample is shown below.

Unmarked sample passage:

وفى السنة الرابعة غزا غزوة نعسر، وهو سلطان موش. ومشى معه السيّد المبارك، مور صالح جور، فأمره ان يجعلها جهاداً في سبيل الله. فلم يخالفه في ذلك وبيّن جميع احكام الجهاد. فطلب امير المومنين، اسكيا الحاج محمد، من السيّد المذكور ان يكون مرسولاً بينه وبين سلطان موش. فقبل، ووصل اليه في بلده وبلّغه رسالة اسكيا في الدخول في الاسلام. فقال له حتى يشاور اباءه الذين في الاخرة. فمشى الى بيت صنمهم مع وزرائه. ومشى هو معهم لينظر كيف يشاور الاموات. فعملوا ما يعملون من عوائدهم في صدقاتهم، فظهر لهم شيخ كبير. فلمّا راوه سجدوا له، واخبره الخبر. فتكلّم لهم بلسانهم، وقال لا اقبل لكم ذلك ابداً. بل تقاتلونه حتّى تفنوا عن اخركم او يفنوا عن

اخرهم. فقال نعسر للسيّد المبارك ارجع اليه وقل له ما بيننا وبينه الّا الحرب والقتال. ثمّ قال لذلك الشخص الذى ظهر فى صورة الشيخ بعد ما خرج الناس من ذلك البيت: سالتك بالله العظيم، من انت؟ فقال: انا ابليس، اغويهم لكي يموتوا على الكفر. فرجع الى الامير اسكيا الحاج محمد واخبره بجميع ما جرا. فقال: عليك الان بالقتال فيهم. فقاتلهم، وقتل رجالهم، وخرب ارضهم وديارهم، وسبا ذراريهم. فكلّ من اتى في هذه السبى من رجال ونساء صاروا مباركين. ولم يكن في هذا الاقليم جهاد في سبيل الله الّا هذه الغزوة وحدها.

Sample passage disambiguated using CAMeL Disambiguation:

وَفَى السَنَةِ الرابِعَةِ غَزا غَزْوَةٌ نَعْسُر ، وَهُوَ سُلْطان موش . وَمَشَى مَعَهُ السَيِّدُ المُبارَكِ ، مُور صالِح جَوْر ، فَأَمْره أَنَّ يَجْعَلها جِهاداً فِي سَبِيلِ الله . فَلَم يُخالِفه فِي ذٰلِكَ وَبَيَّنَ جَمِيعِ أَحْكامِ الجِهادِ . فَطَلَبَ أَمِيرُ المومنين ، اسكيا الحاجّ مُحَمَّد ، مِن السَيِّدُ المَذْكُورِ أَنَّ يَكُون مرسولا بَيِّنَهُ وَبَيِّنَ سُلْطان موش . فَقِبَلَ ، وَوَصَلَ إِلَيهِ فِي بَلَدِهِ وَبَلَغَهُ رِسالَةً اسكيا فِي الدُخُولِ فِي الإِسْلامِ . فَقالَ لَهُ حَتَّى يُشاوِر آباءه الَّذِينَ فِي الآخِرَة . فَمَشَى إِلَى بَيْت صَنَمهم مَعَ وُزَرائِهِ . وَمَشَى هُوَ مَعَهُم لِيَنْظُر كَيْفَ يُشاوِر الأَمْوات . فَعَمِلُوا ما يَعْمَلُونَ مِن عَوائِدهم فِي صَدَقاتهم ، فَظُهْر لَهُم شَيْخُ كَبِيرٍ . فَلَمّا رَأَوْهُ سَجَدُوا لَهُ ، وَأُخْبِره الخُبَر . فَتَكَلَّمَ لَهُم بِلِسانهم ، وَقالَ لا أَقْبَلَ لَكُم ذٰلِكَ أَبَداً . بَل تُقاتِلُونَهُ حَتَّى تَفْنَ◌ُوا عَن آخَرِكُم أَو يَفْنَ◌ُوا عَن آخِرُهُم . فَقالَ نَعْسُر لِلسَيِّدِ المُبارَكِ أَرْجِع إِلَيهِ وَقُلْ لَهُ ما بَيِّنَنا وَبَيِّنَهُ إِلّا الحَرْبِ وَالقِتالِ . ثُمَّ قالَ لِذٰلِكَ الشَخْصُ الَّذِي ظُهْرِ فِي صُورَةِ الشَيْخ بُعْدَ ما خَرَجَ الناسِ مِن ذٰلِكَ البَيْتِ : سَأَلْتُكَ بِاللهِ العَظِيم ، مِن أَنْتَ ؟ فَقالَ : أَنا إِبْلِيس ، أَغْوِيهم لِكَي يَمُوتُوا عَلَى الكُفْرِ . فَرَجَعَ إِلَى الأَمِيرِ اسكيا الحاجّ مُحَمَّد وَأُخْبِره بِجَمِيعِ ما جَرّاً . فَقالَ : عَلَيكَ الآنَ بِالقِتالِ فِيهِم . فَقاتِلهم ، وَقُتِلَ رِجالهم ، وَخَرَّبَ أَرْضِهِم وَدِيارِهِم ، وَسَبَأ ذَرارِيّهم . فَكُلُّ مِن أَتَى فِي هٰذِهِ السَبْي مِن رِجالِ وَنِساءِ صارُوا مُبارَكَيْنِ . وَلَم يَكُن فِي هٰذا الإِقْلِيمِ جِهاد فِي سَبِيلِ الله إِلّا هٰذِهِ الغَزْوَة وَحْدَها .

Conclusion

This paper provides a summary of the various tools applied to selected passages from the Timbuktu Chronicles, as wells as some of the issues that arose. There are of course several potential next steps that may be considered, including applying the above-mentioned tools to the entire texts as opposed to just samples of them. One step is, of course, training OCR models to improve the Arabic text files and make them as similar to the original as possible. Another step could be to integrate geo-tagging using the NER labels. NER already highlights the geographical and geopolitical entities, so it would be useful to link those entities to a map to visualize them and the interactions between them. For example, the geopolitical locations labelled in the chronicles could be integrated with a project such as al-Ṯurayyā, a web-based gazetteer that geospatially models the early Islamic world, to include outskirt of what is normally thought of the "Islamic world" in the scope. Another possible step could be to display various samples of the texts on a web page or application with capabilities for user interaction and tool selections – for example, integrating with a project such as the Scaife Viewer.

Bibliography

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. https://spacy.io/

Leung, Kenneth. *Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER).* June 2021. https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510 and https://github.com/kennethleungty/OCR-Metrics-CER-WER/

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. https://stanfordnlp.github.io/stanza/

Romanov, Maxim and Seydi, Masoumeh. *Al-Thurayya Project.* https://althurayya.github.io/

Other tools, resources, and references:

CAMeL Tools: https://github.com/CAMeL-Lab/camel_tools

Scaife Viewer: https://scaife.perseus.org/

Displacy: https://spacy.io/usage/visualizers

Tarikh al-Fattash in Arabic:
https://babel.hathitrust.org/cgi/pt?id=hvd.hndhq3&view=1up&seq=15&skin=2021

Tarikh al-Fattash in French: https://hdl.handle.net/2027/wu.89012185690

Tarikh as-Soudan in Arabic:
https://babel.hathitrust.org/cgi/pt?id=hvd.32044035034420&view=1up&seq=81&skin=2021

Tarikh as-Soudan in French: https://hdl.handle.net/2027/hvd.32044019989060

DeepL: https://www.deepl.com/translator