

# Increasing the Accessibility of the Autobiography of Omar ibn Said through Translation Alignment

# Joseph C. Hilleary

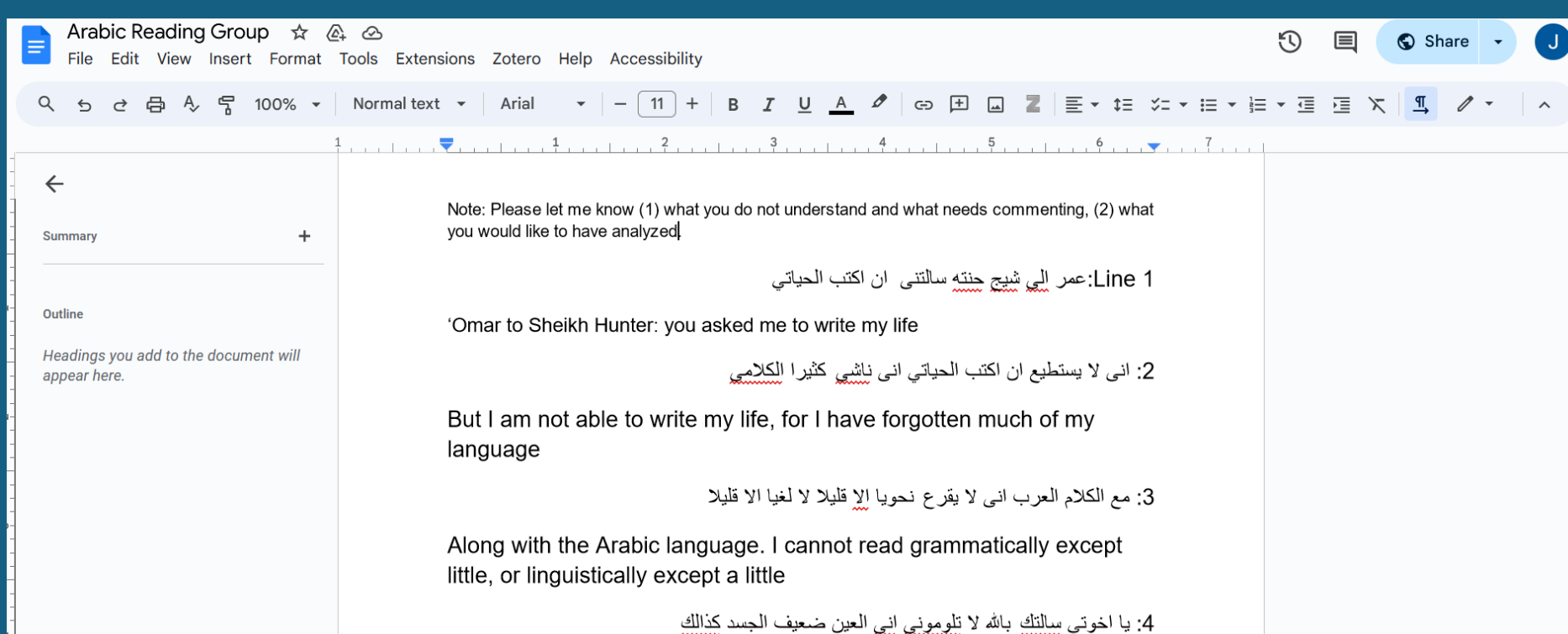
Tufts University, Department of Computer Science

# Introduction

In 1807, Omar ibn Said, a West African man educated in the Islamic tradition, was captured, enslaved, and brought to the Carolinas. Roughly two decades later, he produced a first-hand account of his life written in Arabic. The original manuscript forms the centerpiece of a collection of now digitized documents held by the Library of Congress. This work is notable because unlike other contemporaneous slave accounts written in English, it was not subject to editing by the slave-holder ibn Said lived with and can be viewed as a more faithful version of his experience. In addition, it demonstrates the strong traditions of writing and education that existed in the parts of Africa that were raided during the trans-Atlantic slave trade.

In order to make this important document more accessible, particularly in a secondary classroom setting where many students may not have the requisite language skills to read the original Arabic, we have produced a lightly annotated digital edition of the text aligned to both an original translation and a transliterated version of the Arabic. The goal of this work is to give readers with limited Arabic the confidence to engage with a primary source and to elevate the interesting components of the manuscript that disappear when it is only read in translation.

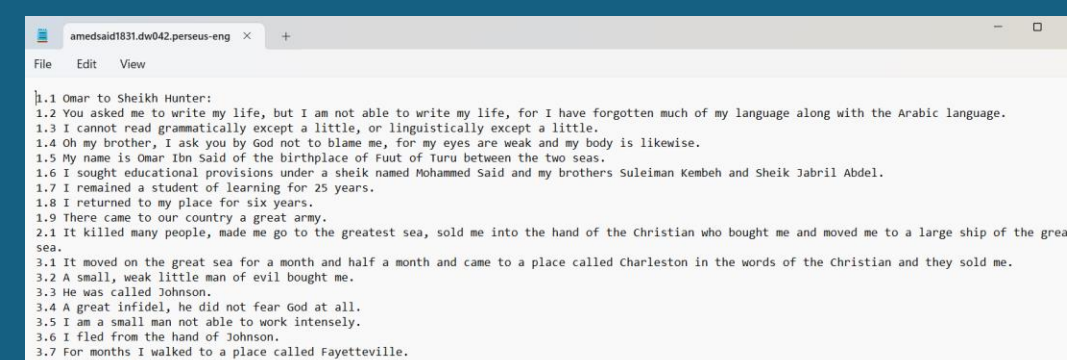
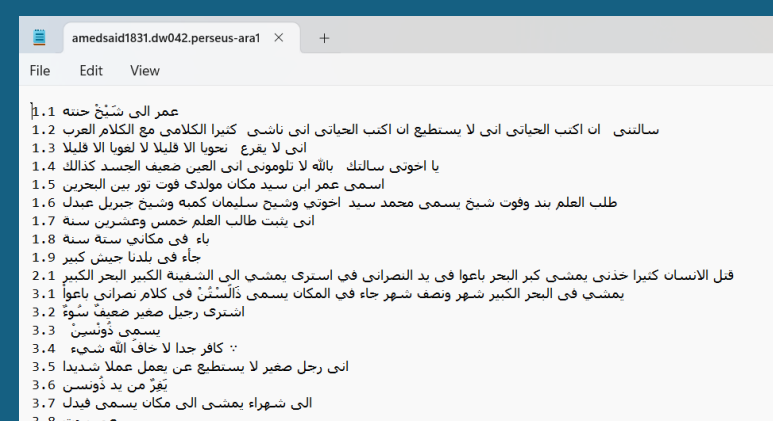
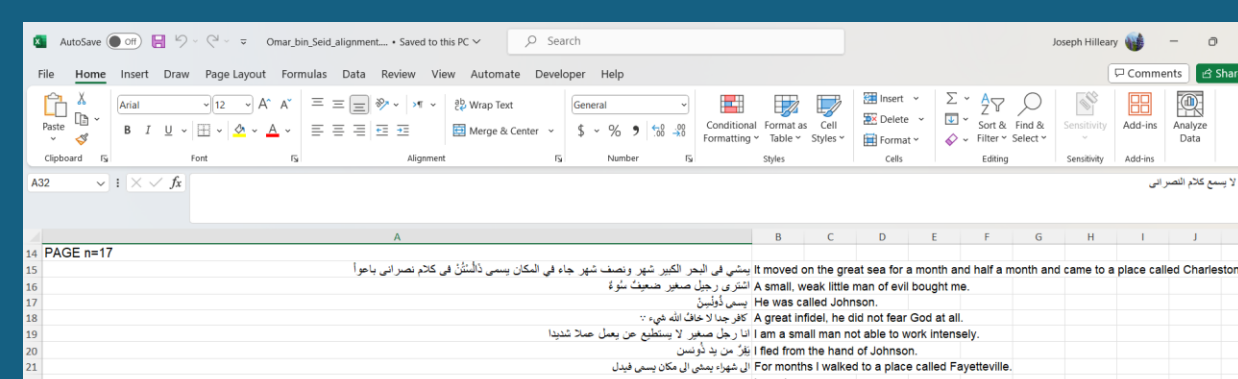
# Collaborative Transcription & Translation



While the contents of the manuscript are in the public domain, some of the existing English translations are not. During the fall of 2022, as part of a reading group for a course on West African Classics led by Prof. Gregory Crane at Tufts University, several students with very levels of Arabic fluency produced a new translation of biographical section of the manuscript, focusing on literal adherence to the original text.

Translation by  
Ingrid Barnes, Klara Chura, Gio DiRusso, Joseph Hilleary, Nole Wade

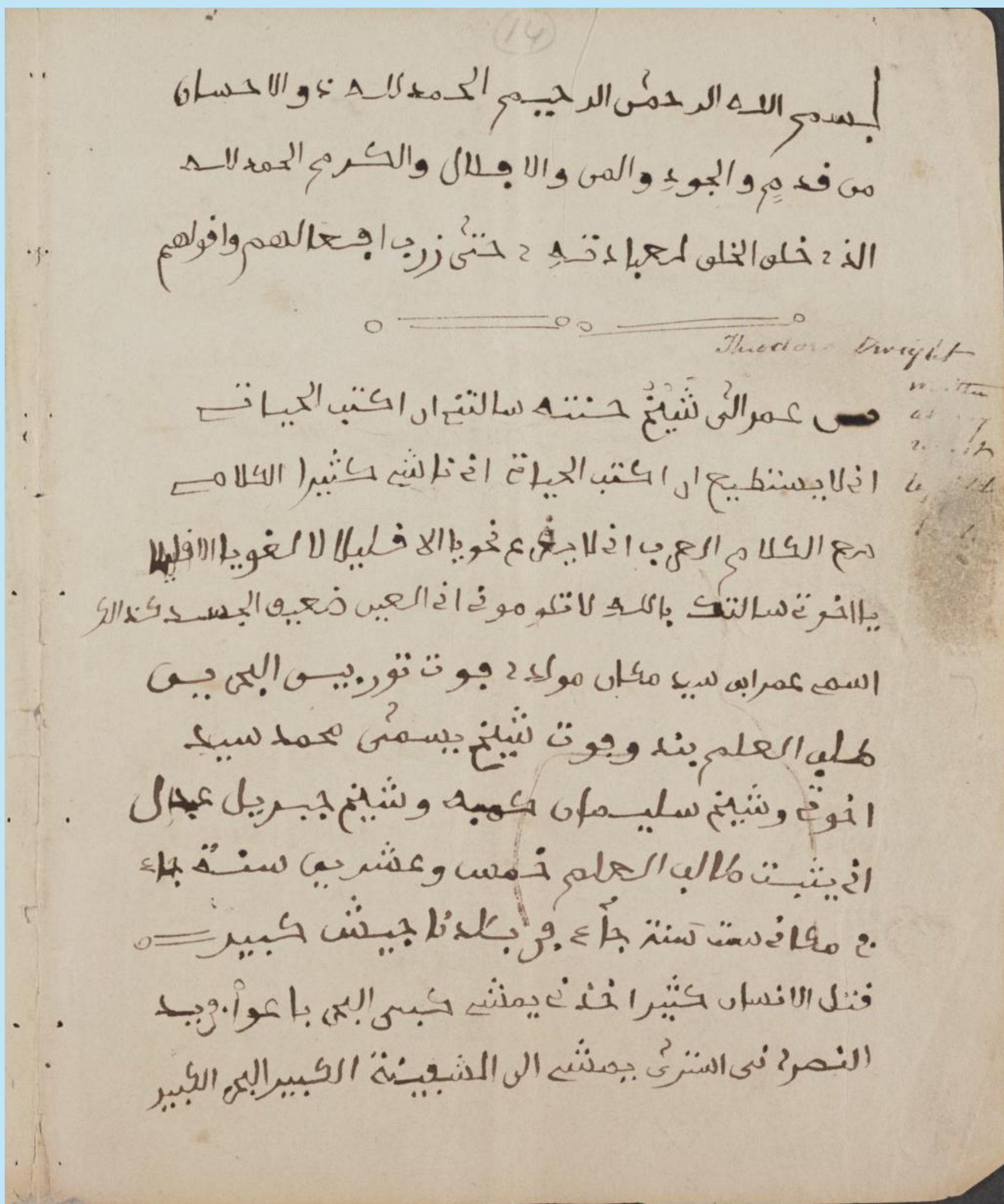
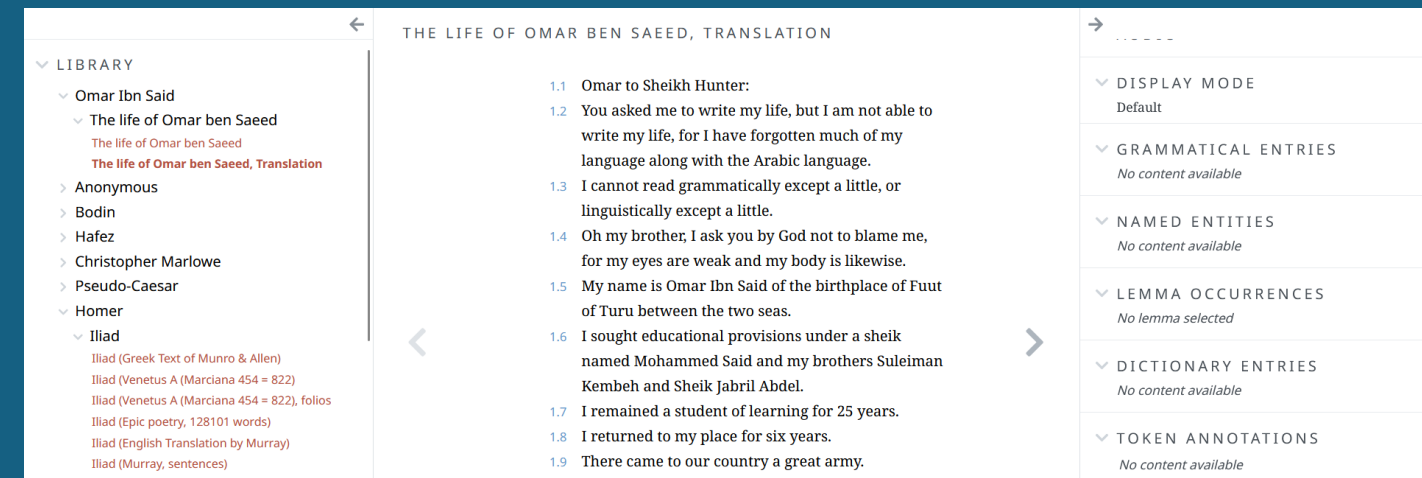
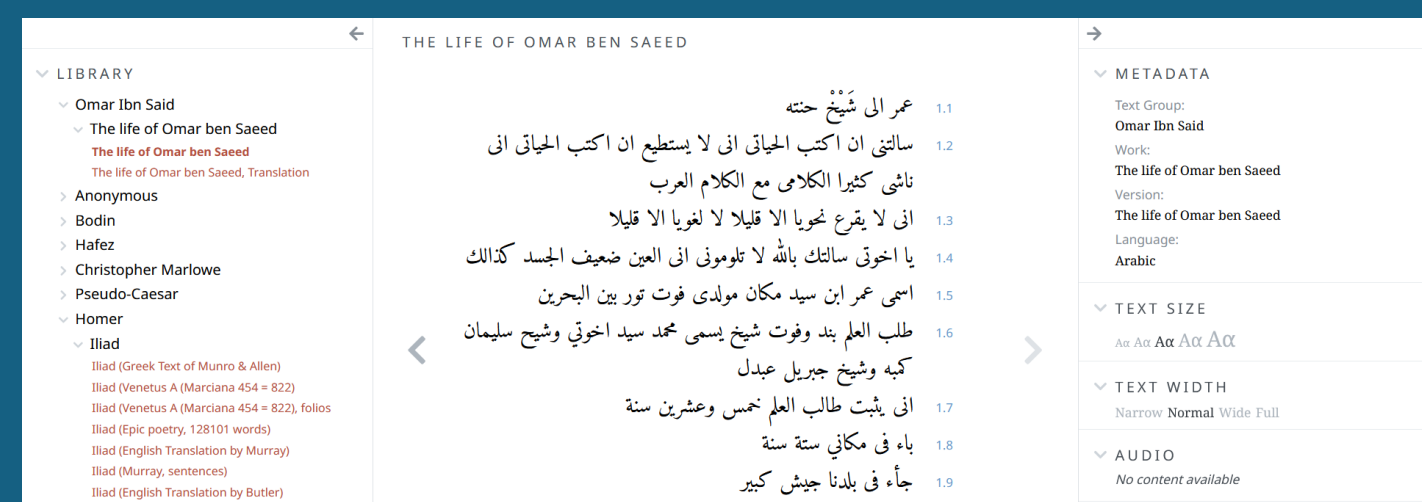
# Sentence Alignment & Text File Prep



## Beyond Translation & Future Work

Ultimately, our edition will be published in Beyond Translation, the 6<sup>th</sup> generation of the Perseus Digital Library. Built on top of the Scaife Viewer, the Beyond Translation reading environment will allow for interactive engagement with our annotations. At the same time, the inclusion of the Ibn Saïd autobiography in the Perseus collection, already widely used in scholastic settings, will automatically put it at the fingertips of educators across the country.

Beyond Translation also supports many additional forms of text enrichment. Future work on this document could include maps of geospatial references, lists of named entities with corresponding links to existing knowledge bases, IIIF links to the images of the original manuscript, tracking of lemma usage, and/or an integrated Arabic grammar.



## Annotations

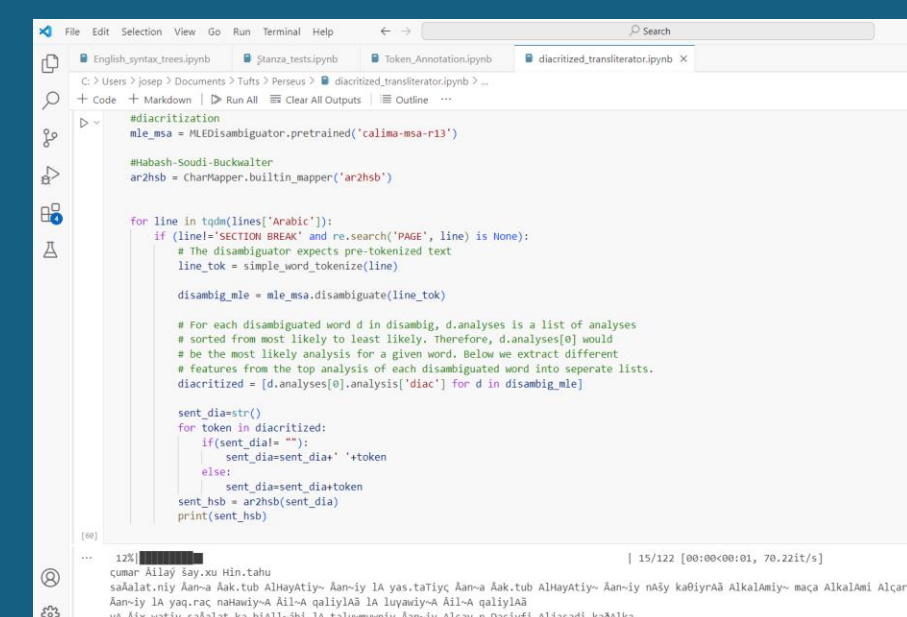
	A	B	C
1	token_id	token	alt_text
2	1.1.1	عمر	
3	1.1.2	التي	
4	1.1.3	بنيت	
5	1.1.4	هذه	
6	1.2.1	سأنتي	التي
7	1.2.2	ان	
8	1.2.3	الكتب	
9	1.2.4	الحياتي	حياتي
10	1.2.5	الي	ها
11	1.2.6	لا	
12	1.2.7	يستطيع	يستطيع
13	1.2.8	ان	
14	1.2.9	الكتب	
15	1.2.10	الحياتي	حياتي
16	1.2.11	الي	ها
17	1.2.12	نأسي	نأسي
18	1.2.13	كثيرا	
19	1.2.14	الكلامي	لامي
20	1.2.15	مع	
21	1.2.16	الكلام	
22	1.2.17	العرب	
23	1.3.1	الي	

## Token annotations

Omar ibn Said was writing in a second language after years of minimal use. Throughout the manuscript he uses irregular spellings and orthography. We wanted to both correct these to improve legibility, while preserving the original decisions which tell their own story. Many of ibn Said's "mistakes" were recognizable to Arabic learners as things they struggled with as well, such as the mixing up  $\tilde{r}$  and  $\tilde{l}$ .

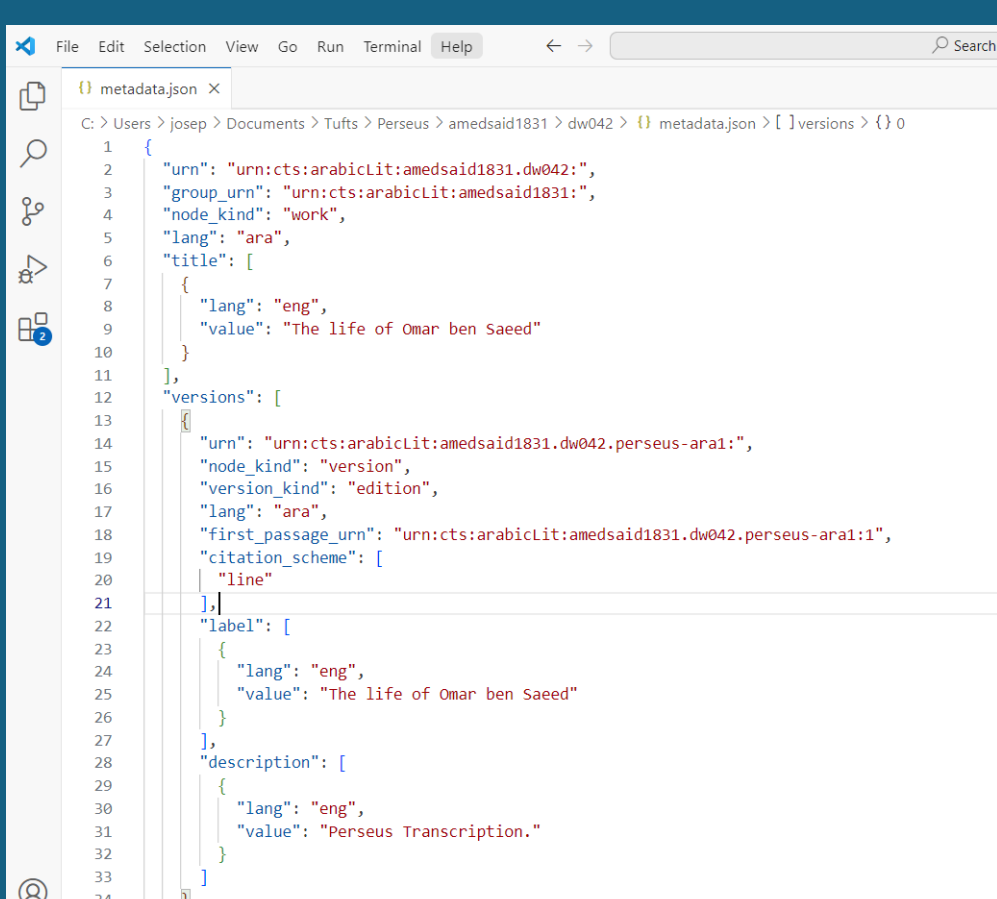
## Transliteration

Many in the intended audience for our edition may be unfamiliar with the Arabic script. Using the Python library CAMEL tools developed by Obeid et al. at NYU Abu Dhabi, we conducted morphosyntactic analysis of our transcription in order to restore the diacritics. Then, using this enriched version, we produced a Buckwalter transcription of the document.



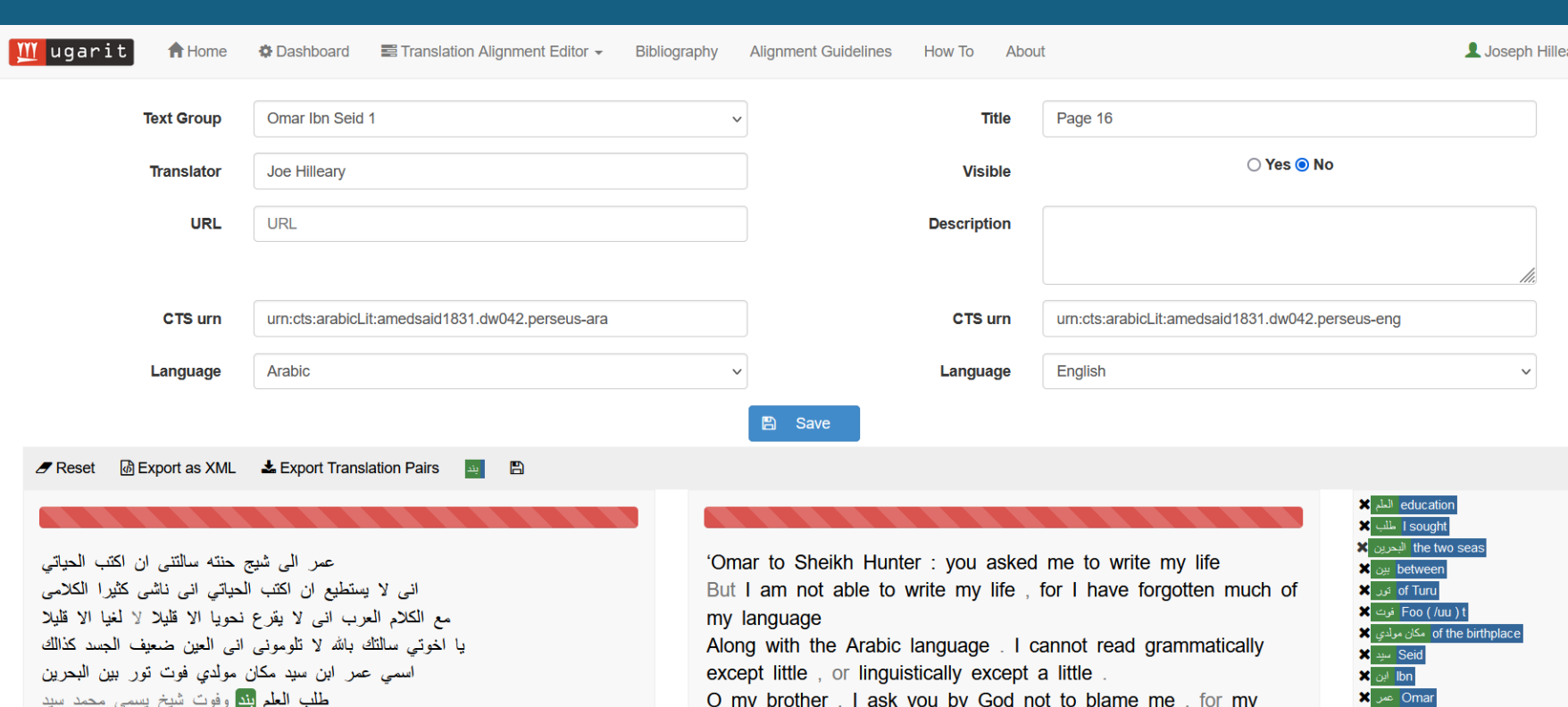
## CTS Metadata

In order to facilitate the digital service of our edition, we minted metadata for the Ibn Said Autobiography following the Canonical Text Services protocol (Blackwell and Smith). We based the new URN on the existing metadata categorizing the manuscript within the Library of Congress collections to maximize discoverability and compatibility.



# Word Alignment

One of the most effective tools for readers engaging with materials in unfamiliar languages is a word-level aligned translation, which facilitates understanding of both vocabulary and syntax. Using the UGARIT online alignment editor created by Yousef et al. at Universität Leipzig, we manually aligned each word of our translation with the original Arabic. This alignment data was then exported to a flat file which could be converted into the appropriate format for our intended digital reading environment.



## Treebanking

Syntax trees provide insight into sentence structure, helping readers in an unfamiliar language follow how words relate to one another. Using Stanza, a project by Qi et al. at the Stanford NLP Group, we produced trees for the translation in the Universal Dependencies system. At the time of the work, the existing UD parsers were ineffective for Arabic, so the transcription was manually treebanked using the Palmyra editor by Taji and Habash (also NYU Abu Dhabi).

