

Applying Computational Text Analysis to *One More Voice*'s “Corpus of Africa-Centered Literary Works, 1830-1930”

Caitlin Matheis (University of Nebraska, 2021)

Table of Contents

- I. Project Overview
- II. Corpus Planning and Building
- III. Metadata Standardization
- IV. Initial Voyant Experimentation
- V. Learning Python
- VI. Conclusion

I. Project Overview

The goals of this computational text analysis project were to 1) create a corpus of nineteenth and early twentieth-century Africa-centered literary works for [One More Voice](#) and 2) perform computational text analysis on the corpus.

Using works that were recommended by scholars in the field of Victorian-era, African-centered literature, we identified existing digitized editions of the earliest editions of each work using [Internet Archive](#), [HathiTrust](#), and other such sources. From those digital sources, I created plain text (TXT) files so that they could be uploaded and analyzed. I then collected metadata for each work and stored it in [Zotero](#), then a spreadsheet.

I performed computational text analysis using [Voyant](#), an online tool for text analysis. I informally recorded observations and analysis each time that Voyant was used to examine the corpus in order to reflect on results and ask questions that would inform further research. I also researched the [Python](#) programming language to understand how it might extend the computational analysis carried out through Voyant.

Examining the corpus compiled for *One More Voice* in Voyant suggested information about trends and themes throughout the corpus. The results from several tools available in Voyant, including topic modeling tools and word frequency tools, provided insight who was speaking in the texts, the linguistic makeup of the corpus as a whole, and what subjects were primarily discussed throughout the corpus. Research suggested that primarily British, male, characters were speaking in the texts, that religion was discussed with surprising infrequency, and that the language of colonization is present in the corpus whether explicitly discussed or not. While men seem to be speaking across the majority of the corpus, British male names appeared to be used more frequently in British-authored texts.

II. Corpus Planning and Building

A. Results

As of this writing, the *One More Voice* “[Corpus of Africa-Centered Literary Works, 1830-1930](#)“ contains 62 complete Victorian-era literary texts from British and African authors.¹ Originally compiled with the purpose of beginning a computational text analysis project for *One More Voice*, the corpus is now [available](#) from the *One More Voice* GitHub repo for consultation, research, and analysis purposes.

The texts of the corpus are primarily nineteenth- and early twentieth-century fiction and non-fiction travel narratives that take place in Sub-Saharan Africa. There are also several works of folklore, fables, nursery rhymes, and grammar. There are currently a close to equal number of works of fiction and non-fiction in the corpus. While the corpus contains equal representation of African and British authors, most of the authors in the corpus are male, though there are four female authors with texts in the corpus: Mary Prince, May French-Sheldon, and Mary Henrietta Kingsley, who were the sole authors of their works; and Lucy Catherine Lloyd, who co-authored *Specimens of Bushman Folklore* with W.H.I. Bleek.

Several sources were consulted for the compilation of the corpus, including works that were already cited on *One More Voice*'s “[Book-Length Works](#)“ page. Several works were also recommended by Dr. Adrian S. Wisnicki (University of Nebraska-Lincoln), the lead developer of *One More Voice* project. In addition, several fictional works were also recommended by Justin Livingstone (Queen's University Belfast), a *One More Voice* contributor. Finally, peer review of the corpus as a whole and further recommendations were provided by Ng'ang'a Wahu-Mũchiri (University of Nebraska-Lincoln). The text of the earliest edition – in many case the first edition – of each work was found on the *Internet Archive* or *HathiTrust*.

B. Process

[Zotero](#) and a spreadsheet were used to store and organize the metadata for the project as the *One More Voice* corpus was compiled. In addition, one of the goals in building the corpus was to locate the earliest edition of the work that has been digitized; such works were mostly found through the [Internet Archive](#) and [Hathi Trust](#). In situations where there was not a digitized first edition of a text, the earliest edition that could be found was added to the corpus. The corpus includes works of fiction and non-fiction that take place in Sub-Saharan Africa, from both African and British authors. While many of the texts in the corpus are longer works already cited on the *One More Voice* website, there are also British-authored adventure novels, travel

¹ For the purposes of determining this number, we counted two-volume works as two separate texts and, indeed, in the corpus each volume of a multi-volume work is presented in a separate text file.

narratives, and other works that were recommended by several *One More Voice* contributors and through peer review (see prior subsection).

The corpus was built with the intention of having equal representation between British-authored and African-authored works, with the goal being that this corpus should not silence the writing of the African authors. More British-authored accounts of imperialism in and the colonization of regions in Africa during the Victorian era have survived than African-authored narratives or accounts. But those are not grounds for including more British-authored texts and inherently making the African-authored texts seem less significant, thereby potentially producing data that only continues to uphold colonization and white supremacy.

The first works compiled into the corpus were from *One More Voice*'s "[Book-Length Works](#)" page. Only texts on the site that already had links to full text versions were included. Dr. Adrian S. Wisnicki, the lead developer for *One More Voice* also recommended several British-authored Victorian-era texts to be included in the corpus.

After the works were located and saved, the metadata for each text was standardized in Zotero. After this initial compilation, there were significantly less fiction texts than non-fiction texts. Justin Livingstone, a *One More Voice* contributor, whose scholarly work focuses on Victorian adventure novels, recommended the following works, which were added to the corpus:

- H. M. Stanley, *My Kalulu: Prince, King, and Slave: A Story of Central Africa* (Sampson, Low, Marston, 1873)
- Verney Lovett Cameron, *Jack Hooper: His Adventures at Sea and in South Africa* (T. Nelson, 1887)
- Samuel Baker, *Cast Up by the Sea* (Macmillan, 1868)
- R. M. Ballantyne, *Black Ivory: A Tale of Adventure Among the Slavers of East Africa* (T. Nelson, 1873)
- R. M. Ballantyne, *The Gorilla Hunters: A Tale of the Wilds of Africa* (T. Nelson, 1861)
- W. H. G. Kingston, *In the Wilds of Africa: A Tale for Boys* (T. Nelson, 1872)
- W. H. G. Kingston, *The Two Supercargoes, or, Adventures in Savage Africa* (Sampson Low, Marston, Searle & Rivington, 1878)
- G. A. Henty, *The Young Colonists: A Story of the Zulu and Boer Wars* (Routledge, 1885)
- G. A. Henty, *By Sheer Pluck: A Tale of the Ashanti War* (Blackie, 1884)
- George Manville Fenn, *Off to the Wilds: Being the Adventures of Two Brothers* (Sampson Low, Marston, Searle & Rivington, 1881)
- George Manville Fenn, *A Dash from Diamond City* (Ernest Nister, 1901)
- John Buchan, *A Lodge in the Wilderness* (Blackwood & Sons, 1906)
- John Buchan, *Prester John* (T. Nelson, 1910)

Ng'ang'a Wahu-Mũchiri, another *One More Voice* contributor, peer reviewed the corpus and recommended that works by Samuel Ajayi Crowther and Thomas Birch Freeman be added. Based on this recommendation, the following works were added to the corpus:

- Samuel Ajayi Crowther and John Christopher Taylor, *The Gospel on the Banks of the Niger. Journals and Notices of the Native Missionaries Accompanying the Niger*

Expedition of 1857-1859 (Church Missionary House, Seeley, Jackson and Halliday, 1859)

- Samuel Ajayi Crowther, *A Vocabulary of the Yoruba Language* (Seeleys, 1852)
- Samuel Ajayi Crowther, *Journal of an Expedition Up the Niger and Tshadda Rivers Undertaken by Macgregor Laird in Connection with the British Government in 1854* (Church Missionary House, Seeley, Jackson and Halliday, 1855)
- Samuel Ajayi Crowther, *A Charge Delivered on the Banks of the River Niger in West Africa* (Seeley, Jackson & Halliday, 1866)
- Thomas Birch Freeman, *Missions in Western Africa: Including Mr. Freeman's Visit to Ashantee* (Carlton & Porter, 1842)
- Thomas Birch Freeman, *Journal of Two Visits to the Kingdom of Ashanti, in Western Africa* (John Mason, 1843).

C. Summary

The initial corpus for the *One More Voice* computational text analysis project contains 62 complete texts by British-authored and African-authored. Many of the texts were taken from the *One More Voice* website, while several were recommended by Victorian-era literary scholars. Further works may be added to the corpus in the future.

III. Metadata Standardization

A. Results

Using Zotero and a spreadsheet, the following metadata was recorded for the sake of organization and for informing an analysis of the corpus as a whole:

- TXT File Name
- Short Title
- Author 1
- Author 2
- Publication Date
- Publication Title
- Volume
- Place 1
- Place 2
- Place 3
- Place 4
- Place 5
- Publisher 1
- Publisher 2
- Publisher 3

- Publisher 4
- Item Type
- Genre
- Author 1 Origin
- Author 2 Origin
- Author 1 Gender
- Author 2 Gender
- TXT File Word Count
- TXT Source
- TXT Source URL

Each plain text (TXT) file was downloaded from the corresponding digital version of each work. The following statement was also added to the beginning of each TXT file:

About This File

This file is part of a corpus compiled by Caitlin Matheis (University of Nebraska-Lincoln) for a 2021 computational text analysis project for the *One More Voice* recovery project (onemorevoice.org). This corpus encompasses a wide range of plain text (TXT) files containing nineteenth- and early twentieth-century literary works centered on Africa by several dozen different British and African writers. Questions about the corpus and project can be directed to Caitlin Matheis (cmatheis2@huskers.unl.edu), the creator of the corpus, and Adrian S. Wisnicki (awisnicki@unl.edu), lead developer for *One More Voice*.

After this statement, a URL to the text source was included for each work.

At this stage of the project, very few alterations were made to the downloaded texts. Any text that preceded the title page of the scanned work was eliminated from the file, as this text tended to be extraneous text introduced through the OCR process; no deletions were made later in the works. Afterwards, the title page text of each work was cleaned up, so that each work was clearly identified in the contents of each files, in addition to the file name.

B. Process

Most of the corpus metadata was individually tracked in [Zotero](#). While several of Zotero's metadata fields were used to record information for each work and volume in the corpus, other metadata (such as the gender and nationality of the author or the work and the genre of the text) was also recorded because of the insight it would likely provide once computational text analysis was performed.

Many of the titles of the works in the corpus are quite long. While these could have been abbreviated during the metadata standardization process, the entire original titles were recorded; the length and structure of the titles are characteristic of the genre of texts included in the corpus.

Additionally, several texts in the corpus have the same author. Though it was clear in these editions that they were written by the same author, many had originally been published with slight variations in the name. For example, one work was listed as being written by Samuel Crowther, while another was listed as being written by Samuel Ajayi Crowther. In these situations, the longest published form of the author's name was recorded on all of their works in Zotero and the metadata spreadsheet.

Several of the works contained multiple volumes. Because each volume of a work was uploaded onto the [Internet Archive](#) and [Hathi Trust](#) separately – as opposed to putting each physical volume into one digitized edition – a separate TXT file was created and cited in Zotero for each individual volume of a work.

Many of the works had multiple publishers and publication places. Each of these was recorded in the metadata spreadsheet.

The corpus plain text files were generated from OCR scanning of PDFs of the editions that were chosen to be included in the corpus. It was originally hoped that [Project Gutenberg](#) would be a useful resources for cleaned up versions of each work in the corpus in order to create a smaller margin of error in the computational text analysis process. However, given that each work was not on *Project Gutenberg's* site, that many of the editions from which the digital text was taken was not stated at the beginning of each web page, and that the changes that were made to the texts in the digitization process were not immediately obvious, it was ultimately decided that the OCR text from the source page would be used to create the corpus.

Once each TXT file was downloaded, the file name and word count of the file was recorded with the other metadata.

While the entirety of each plain text file could not be cleaned up, anything scanned by OCR before the frontispiece of the text was deleted from the file. In order to clearly identify the work, volume, and the publication information, the title page information was cleaned up to be represented as it appears in the original work.

Finally, a general statement was included at the beginning of each TXT file to indicate that it was created as part of the *One More Voice* corpus.

C. Summary

Several pieces of key metadata about each of the works were gathered, standardized, and recorded in order to ensure that a thorough bibliography of the corpus was maintained. The information that was chosen to be gathered was intended to aid in the computational text analysis process while also remaining true to the genre of the works themselves.

The plain text (TXT) files were gathered for each work. The beginning of each file for each work of the corpus was amended with the intention of clearly identifying the work in the file, the file source, and information on the corpus and *One More Voice*.

IV. Initial Voyant Experimentation

[Stable version](#) to *One More Voice's* Corpus in Voyant.

A. Initial Results

[Voyant](#) was the primary tool used for researching *One More Voice*'s corpus. The user-friendly, web-based tool is accessible for beginners in computational text analysis, especially because it contains a wide range of different analysis and visualization tools.

Despite trying to enter the corpus exploration with as little expectation or bias as possible, there were certain results that were initially anticipated. Because the text files were in fairly raw form, there was also an assumed margin of error in the results. In addition, many place names or references to places were expected in the initial Voyant results due to the genre of texts that were being analyzed.

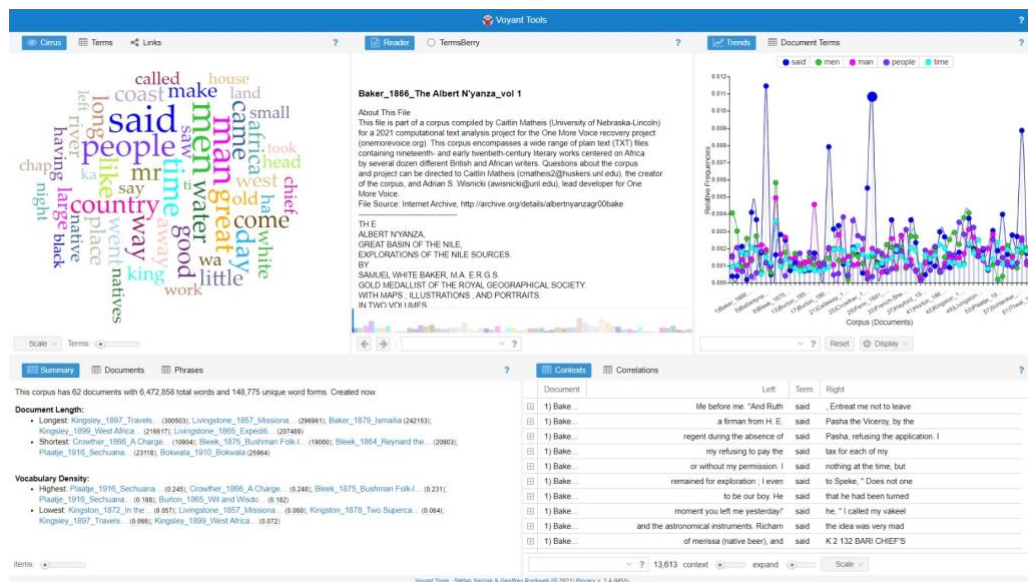


Fig. 1. Voyant home screen after uploading the One More Voice corpus (file name: omvData_0018)

In Fig. 1, the first Voyant screen viewable after uploading the corpus, there is both information that doesn't seem particularly surprising and information that is intriguing. Research focus was first placed on the word cloud in the top left corner. The word cloud provided the frequency of individual words in the corpus, ordered from most frequent to least frequent. Knowing the genre and having a general idea of the contents and subjects of the works in the corpus, there were words on this list that were not exactly surprising.

For instance, that “said” is the most frequently used word in the corpus is not particularly surprising: many of the texts incorporated into the corpus are narrative-based, whether fictional or non-fictional.

“Men” and “man,” the second and third most frequent words, while not words that were expected to be seen so high on this list, are also not words that were expected to be used infrequently throughout the corpus.

Additional keywords might be grouped as follows:

Expected and unsurprising words: country, Africa, coast, river, white, native, natives, chief, king, land, black, African, village, lake, home, government, slave

Unexpected but not surprising: left, work, ground, south, water, day, like, went, come, long, place, work, took, make, life, days, house, seen, right, near, new, told, miles, east, taken, look, tree, brought

Surprising Words: chap, little, head, small, know, feet, side, great, good, came, large, old, round, shall, young, half, hand, gave, present, high

Words to look into further:

- **Black, white and Africa/African** are in the top 50 or so words, but nothing referring to English/British/Britain/England, or any other European country (“English” is the first instance and is the 73rd most frequent word).
- In just looking at the top 50 most frequent words, the frequency of positively connotated words, like “**great**” and “**good**” is surprising. It would be particularly interesting to see whose texts those words were most recurring in, and in what context they were used – and who things were “**good**” and “**great**” for. The common use of these words in these texts is not surprising; it is just surprising that they're so high up on the list. “**Like**” may be worth investigating with this as well.
- The frequency of “**old**” is intriguing. And particularly in looking how it is used to talk about differences before and after colonization, or as a word potentially used by the colonizers to encourage colonization. It may be interesting to examine “**make**” and “**took**” in the same kind of context. Maybe even “**seen**”? and “**right**”? “**New**”? “**Taken**”? “**Look**”? “**Like**”? “**Brought**”? “**Came**”?
- “**Said**” is a key structural component to narrative-based works and can indicate who is speaking in the texts.

Because the goal of *One More Voice* is to recover non-European voices from the British imperial archives, further data was prepared to research who is actually speaking throughout these works. This list of the most frequently occurring words that indicate someone is speaking was created in order to aid in completing this research:

- Spok* (for spok, spoken), spoke*
- Say(*), saying, says
- Called, call(*)

- Tell(*), told
- Ask(*), asked
- Cry, cried, (crie*)
- Replied, reply(*), repli(*)
- Observ(*), observed
- Continue(*), continued
- Describ(*), described
- Declared, declar*
- Added, add*
- Exclaimed, exclaim*
- Explained, explain*
- Mentioned, mention*
- Speak*?
- Hear(*) (hear, heard)²
- Answer(*)?
- Think(*), thought
- Considered*?
- Wished*?
- Refused*?

Some research was also completed into the frequency of religious terms throughout the text. While the search indicated that themes of religion were not the most common in the corpus as a whole, there were works where religious terms occurred significantly more frequently compared to other works in the corpus.

Because positively-connotated and negatively-connotated words can tell us how certain subjects were written about, this list of the most frequently occurring positive and negative words in the corpus was generated. Specific category lists can be created in Voyant to explore these specific terms:

Positive:

- Great
- Good
- Kind
- Strong
- Beautiful
- Greater
- Rich
- large

Negative:

- Poor
- hard
- bad

² This may be useful for scholars choosing to study only those who are directly quoted in the texts.

- broken
- impossible
- little
- long
- small
- best

The lists created and presented above served as the initial impression of the contents of the corpus and informed the research that is described in the following sections. Of course, these lists can be used as the subject of research into the corpus or, particularly in the case of the positive and negative lists, to supplement a broader research question.

B. Process

1. Examination of the Corpus Using Voyant

This section of the paper examines several words that could potentially hold significance throughout the corpus, including the use of “take” (“tak*”) and “took” (“took*”), “old,” and “wild.” The words that were chosen to be examined in this analysis were based on the results of the most frequent words in the corpus.

With these lists in mind, results from examining the use of “took*” and “tak*” were examined because of the appropriation of land and people through imperialism and colonization. Searching through the corpus for these words could potentially reveal to what extent the recognition of that appropriation was embedded within the texts. Adding the asterisk makes “took” and “tak” act as prefixes – Voyant's software will look for words that begin with “took” and “tak” but end with any combination of characters to include in the results.

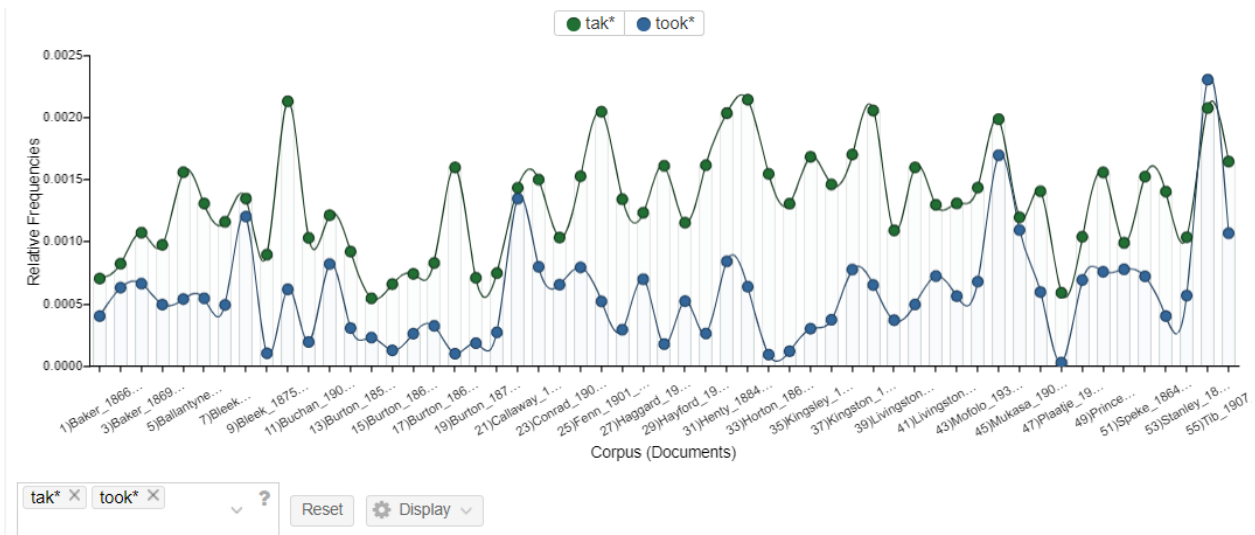


Fig. 2. “Tak*” and “took*” corpus frequency graph (file name: omvData_0009)

When looking at these results, “tak*” is likely to be higher not only because there are more ways that the verb “take” can be conjugated using this prefix than variations of “took*,” but there are also words included in these results that conjugates of the verb “take.” For example, the word “Taka” is included in these results and is used throughout several texts of the corpus to refer to geographical locations such as the Taka Province or the Taka River.

The “Correlations” tool was also used to examine “tak*” and “took*.” The results were arranged the by how frequently “tak*” and “took*” correlate with other words. The closer to one the numerical value of the correlation between two words is, the more likely the two terms are to sync positively, meaning the frequency of one of the words within the text generally increases as the frequency of the other words increases. The closer to negative one the numerical value of the correlation is, the more likely the two words are to sync inversely, meaning the frequency of one of the words generally increases as the frequency of the other word decreases.

At first glance at the results of searching for “tak*” and “took*” in the “Correlations” tool, there doesn't appear to be anything that would, in the “Correlations” column, indicate that the words strongly sync positively or inversely. But the “Significance” column indicates that some of these correlations may be significant.

The closer that the numerical value in the “Significance” column is to 0, the more likely that the correlation between the two words are significant. Generally, significance values less than 0.05 are considered significant. While we have several values that are smaller than 0.05 and approaching zero, Voyant warns that the numbers in this column should be taken with a grain of salt, especially because the data being pulled from the corpus documents entered into Voyant are usually relatively small. Some of the key terms with greater significance that stuck out from the results are shown in Fig. 3.

Term 1	←	→	Term 2	Correlation...	Significanc...
said			tak*	0.38521707	0.0036817...
men			tak*	0.30328444	0.0012785...
man			tak*	0.26437813	0.0006003...
people			tak*	0.23456661	0.0004507...
water			tak*	0.1822133	0.0000454...
like			tak*	0.20706473	0.0000423...
day			tak*	0.17466272	0.0000381...
little			tak*	0.19658709	0.0000328...
time			tak*	0.2570423	0.0000158...
great			tak*	0.24667995	0.0000057...
came			tak*	0.19078012	0.0000022...
good			tak*	0.19104806	7.636991e-7
long			tak*	0.18582761	5.618558e-7

Fig. 3. Results from searching “tak*” in the Voyant correlations tool (file name: omvData_0031)

Term 1	←	→	Term 2	Correlation...	Significanc...
men			took*	0.4143696	0.0000067...
said			took*	0.5859523	0.0000026...
man			took*	0.3569769	0.0000025...
people			took*	0.32671937	7.2300384...
time			took*	0.32082537	5.3117645e-8
like			took*	0.28808954	8.56995e-9
great			took*	0.31121254	7.646616e-9
little			took*	0.2805105	2.1251823...
water			took*	0.2656537	1.928863e-9
day			took*	0.2628458	3.8332515...
good			took*	0.27724528	4.116707e-13

Fig. 4. Results from searching “took*” in the Voyant correlations tool (file name: omvData_0032)

Particularly alarming in these results is the correlation between “tak*” and “took*” with words about life and physical bodies – man, men, people, life, dead, children, and home.

The word “wild” seemed to appear frequently enough throughout several texts to be examined further, particularly the phrase “The Wilds of Africa.” After a quick search in the “Contexts” portion of Voyant for “the wilds of*” (there were 18 results), most of the phrases seem to be “in” or “to” “the wilds of” throughout the corpus. The phrase was followed by “Africa” or by a region in Africa. There were two outliers in these 18 results, that described being in “the wilds of” Florida and Trinidad.

While the contexts search of the wilds produced only 18 results, it would be worth examining whether “wild” was surrounded by positively or negatively connotated words. Based on the initial results, there seems to be several negative words around the use of “wild*,” particularly “brutal” and “hard.”

	Document	Left	Term	Right
⊞	1) Baker_1...	of the past; all is	wild	and brutal, hard [and unfeeling
⊞	1) Baker_1...	Scarcity in View of Plenty—	wild	Duck Shooting—The Crested Crane
⊞	1) Baker_1...	of England's power to rescue	wild	lands from barrenness ; to wrest
⊞	1) Baker_1...	to hardships and endurance in	wild	sports in tropical climates, and
⊞	1) Baker_1...	of Africa I had a	wild	hope, mingled with humility, that
⊞	1) Baker_1...	each rough footstep of the	wild	life before me. "And Ruth
⊞	1) Baker_1...	along its muddy waters in	wild	confusion, bringing a rich harvest
⊞	1) Baker_1...	with the brutality of a	wild	animal. During his administration the
⊞	1) Baker_1...	under such a system. The	wild	speculator borrows upon such terms

wild* x

2,541 context

expand

Scale

Fig. 5. "Wild" contexts in corpus (file name: omvData_0023)

The use of the word "old" in the corpus was then explored. One or two of the documents had "testament" as one of the common following words, and led to questions of how themes of religion are written into the corpus – for example, whether that be religion as a justification for colonization or the discussion of native religions within these texts.

			Term	Count	Trend
⊞	<input type="checkbox"/>	1	god*	1671	
⊞	<input type="checkbox"/>	2	christ*	1161	
⊞	<input type="checkbox"/>	3	relig*	725	

Fig. 6. Religious terms count (file name: omvData_0023)

It is clear from the results shown in Fig. 6 that religiously-connotated words are not used nearly as frequently as other words throughout the corpus.

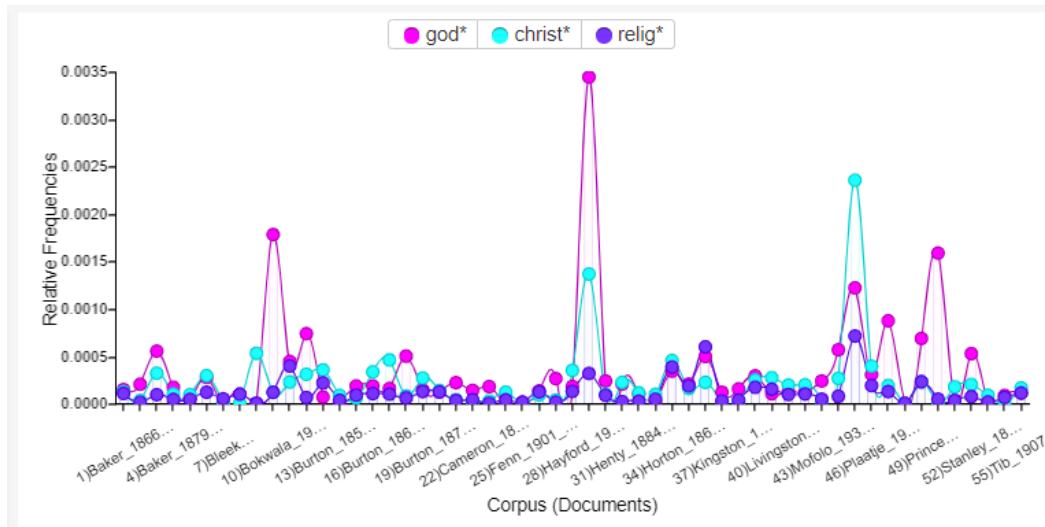


Fig. 7. Religious terms frequency graph (file name: omvData_0024)

The frequency chart of these terms indicates, however, that while it seems that the average frequency appears consistent across several texts, there are a few texts that stand out as using these terms much more frequently than in the others. For the majority of the works, however, religion does not appear to be a topic commonly discussed throughout the works.

2. Topic Modeling

To form a more well-rounded analysis that utilizes several computational methods, the corpus was run through Voyant’s “Topic Modeling” tool. While there are several methods of approaching topic modeling, the “Topic Modeling” tool in Voyant uses Latent Dirichlet Allocation (LDA) and is adapted from David Mimos’ javascript-based tool, [jsLDA](#). Voyant’s “Topic Modeling” tool, like many topic modeling and language modeling tools, needs to be trained in order to produce more accurate results. The tool, by treating the entire corpus as a mixture of topics, is able to produce a series of words called clusters that reveal topics discussed throughout a work or corpus that may not be obvious by just reading the book or series of books. For the *One More Voice* corpus, this means that we can understand what topics are discussed throughout several works of the corpus or what themes are present in the works.

Using this tools, 2000 iterations – the number of times that the corpus was run through the tool in order to produce these results – were created, as shown in Figs. 8 and 9. Looking at the topic clusters of words that appeared after 2000 iterations, it does seem that what appears to be chapter titles. These categories suggest common topics and subjects discussed throughout the corpus, and they also suggest a lot about *how* these subjects are discussed throughout the works. Voyant tells us, using the graphs under the “Scores” category to the right of each cluster, the percentage of each individual text in the corpus that the topic covers in that work; in other words, the more that a cluster appears in a given work, the higher the percentage the results will be.

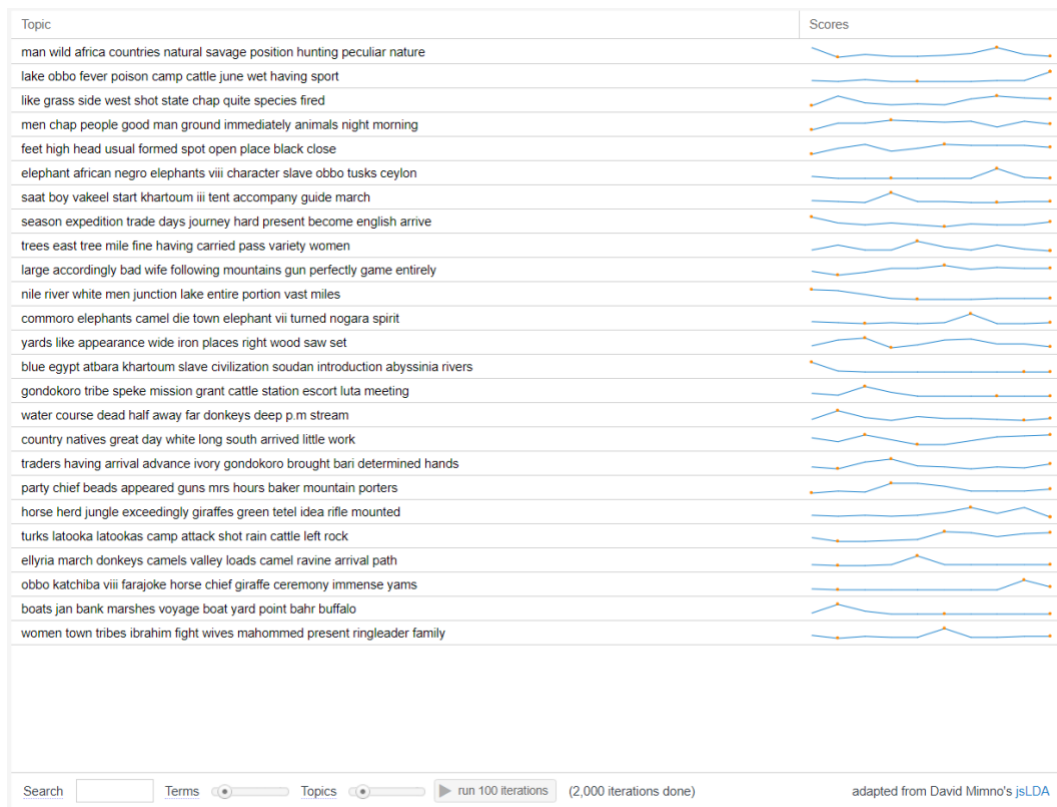


Fig. 8. 2000 Iterations of One More Voice corpus with Voyant's "Topic Modeling" tool (file name: omvData_0025)

Topic
man wild africa countries natural savage position hunting peculiar nature
lake obbo fever poison camp cattle june wet having sport
like grass side west shot state chap quite species fired
men chap people good man ground immediately animals night morning
feet high head usual formed spot open place black close
elephant african negro elephants viii character slave obbo tusks ceylon
saat boy vakeel start khartoum iii tent accompany guide march
season expedition trade days journey hard present become english arrive
trees east tree mile fine having carried pass variety women
large accordingly bad wife following mountains gun perfectly game entirely
nile river white men junction lake entire portion vast miles
commoro elephants camel die town elephant vii turned nogara spirit
yards like appearance wide iron places right wood saw set
blue egypt atbara khartoum slave civilization soudan introduction abyssinia rivers
gondokoro tribe speke mission grant cattle station escort luta meeting
water course dead half away far donkeys deep p.m stream
country natives great day white long south arrived little work
traders having arrival advance ivory gondokoro brought bari determined hands
party chief beads appeared guns mrs hours baker mountain porters
horse herd jungle exceedingly giraffes green tetel idea rifle mounted
turks latooka latookas camp attack shot rain cattle left rock
ellyria march donkeys camels valley loads camel ravine arrival path
obbo katchiba viii farajoke horse chief giraffe ceremony immense yams
boats jan bank marshes voyage boat yard point bahr buffalo
women town tribes ibrahim fight wives mahommed present ringleader family

Fig. 9. Topic modeling clusters after 2000 iterations (file name: omvData_0026)

For example, the first topic listed in Figs. 8 and 9 – “man wild Africa countries natural savage position hunting peculiar nature” – appears to suggest that the countries in Africa and the people that lived in the countries, particularly the problematic assumptions made about those countries and people. Because a significant portion of the works in the *One More Voice* corpus are nonfiction travel narratives by British authors who visited Africa to record their observations and “discoveries,” this topic appears to describe a British perspective – written about in these works or, at least, perpetuated through them – on the lifestyle of the people they encountered.

The topic “season expedition trade days journey hard present become English arrive” suggests us about how expeditions by the English in Africa were discussed – that they were hard, but also motivated by trade.

The topic “elephant African negro elephants viii character slave obbo tusks Ceylon,” while includes “viii,” a chapter heading, tells us that elephants, hunting of elephants, and the comparison of African elephants and Ceylon elephants is commonly discussed throughout these works – and likely a motivation for many of the expeditions that are described throughout the works of the corpus.

3. Investigating Who is Speaking in the Corpus

Because the mission of *One More Voice* is to recover non-European voices from the archive, research was directed towards the fact that “said” is the most frequently appearing word in the corpus.

The top 1000 most frequent words in the corpus, using Voyant's “Terms” tool, were searched to generate a list of “speaking words.” If single verbs appeared on the top 1000, terms that could be searched to include the different conjugations of those verbs in the search were included.

Some of the verbs on this list are likely to appear in conjugations other than those that indicate that someone is speaking. As a result, some words were examined individually to see if they'd be worth including in the overall search to see if they had been used more often to indicate if someone is speaking or not. This search started with “reason*.” While “reason” can be a word that is used to show that someone is speaking, it is much more likely to be used as a noun in the corpus. Voyant's “Contexts” tool indicated that “reason*” is primarily used in its noun form or in its adjective form, “reasonable,” throughout the corpus. Because of this, “reason*” was eliminated from the list of words to be used to search the corpus to find out who is speaking in the text.

Another word that could possibly be eliminated is “question.” Again, a majority the uses of “question” throughout the corpus seem to be the noun form of the word, rather than the verb, so “question” was eliminated from the list as well.

“Start*” was also eliminated because, though it was mostly in the corpus in verb form, it was not directly referring to someone speaking.

Here is the updated list after elimination:

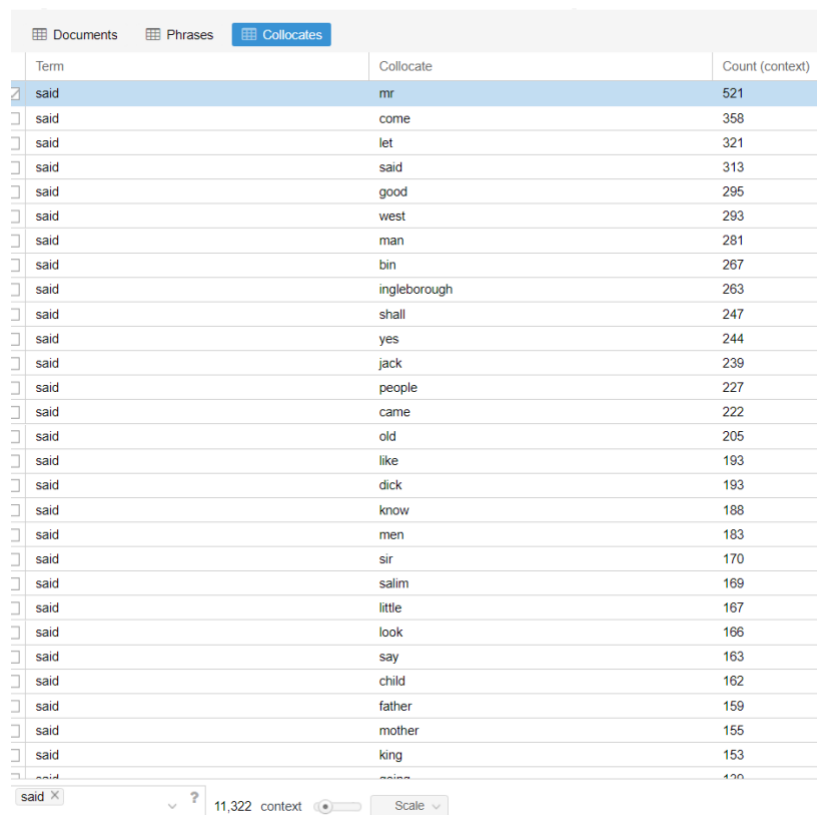
Other possible words to examine in the corpus for who is speaking:

- Spok* (for spok, spoken), spoke*
- Say(*), saying, says
- Called, call(*)
- Tell(*), told
- Ask(*), asked
- Cry, cried, (crie*)
- Replied, reply(*), repli(*)
- Observ(*), observed
- Continue(*), continued
- Describ(*), described
- Declared, declar*
- Added, add*
- Exclaimed, exclaim*
- Explained, explain*
- Mentioned, mention*

- Speak*?
- Hear(*) (hear, heard)³
- Answer(*)?
- Think(*), thought
- Considered*?
- Wished*?
- Refused*?

In addition to compiling the list of verbs that would help research who was speaking within the corpus, I used the “Collocates” tool in Voyant to find which words appear most frequently in context with each other. When users choose this tool in Voyant, the results application automatically shows the top words in the corpus and the words that occur most frequently within a specified distance of five words on either side of that word. Like many of the Voyant's tools, users can choose to look at the results for a specific word.

To begin a further analysis of who is speaking in the texts of the *One More Voice* corpus, I first narrowed the results to show only the words that were collocated with “said,” rather than the most frequent collocations in the corpus as a whole. The first word I used specifically looked at the answer this question was “said.” The top results from the collocates tool for “said” are shown in Fig. 10.



The screenshot shows the Voyant Collocates tool interface. The 'Collocates' tab is selected. The search term is 'said'. The results are displayed in a table with three columns: Term, Collocate, and Count (context). The top results are as follows:

Term	Collocate	Count (context)
said	mr	521
said	come	358
said	let	321
said	said	313
said	good	296
said	west	293
said	man	281
said	bin	267
said	ingleborough	263
said	shall	247
said	yes	244
said	jack	239
said	people	227
said	came	222
said	old	205
said	like	193
said	dick	193
said	know	188
said	men	183
said	sir	170
said	salim	169
said	little	167
said	look	166
said	say	163
said	child	162
said	father	159
said	mother	155
said	king	153

At the bottom of the interface, there is a search bar with 'said' entered, a dropdown menu, and a 'Scale' button.

³ This may be useful for scholars choosing to study only those who are directly quoted in the texts.

Fig. 10. List of words that appear most frequently within 5 words of “said” (file name: omvData_0022)

Most of these initial results seem to imply that throughout the corpus, men are the ones who are speaking with each other in these texts. While this is indicated by the appearance of “king,” “father,” “sir,” “men,” and “father.” When the top results of these terms were character names, many of them were British, male characters from these texts.

However, “said mother” and “said child” stood out in the results of the “Collocates” tool because they weren’t adult men speaking. Interestingly, the results for “said mother” and “said child” more commonly appear in works of folk-lore and nursery tales included in the corpus, like Bleek’s *Reynard the Fox* and Callaway’s *Nursery Tales*. “Said father” also appeared in works that are nursery tales, but also appeared more frequently throughout other texts of the corpus.

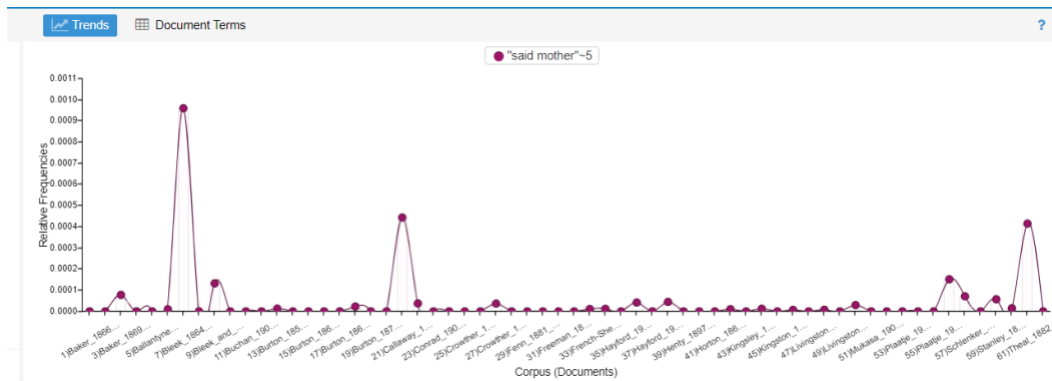


Fig. 11. Frequency graph of how often “said” and “mother” appear within close proximity to each other throughout the corpus (file name: omvData_0027)

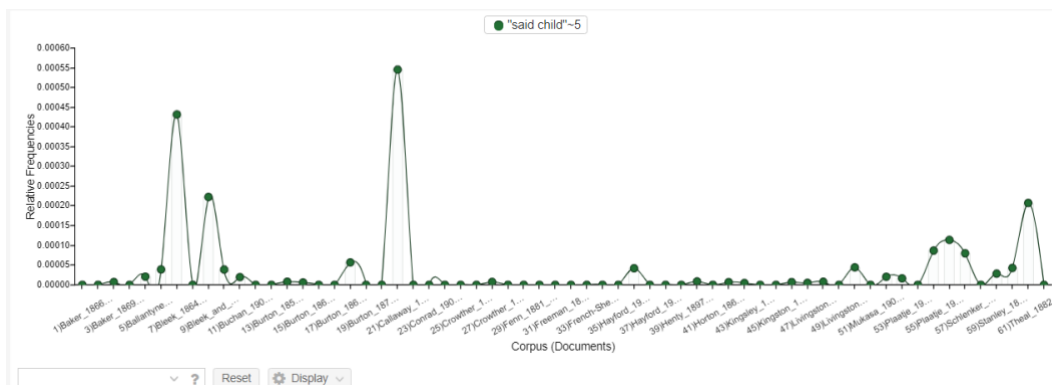


Fig. 12. Frequency graph of how often “said” and “child” appear within close proximity to each other throughout the corpus (file name: omvData_0028)

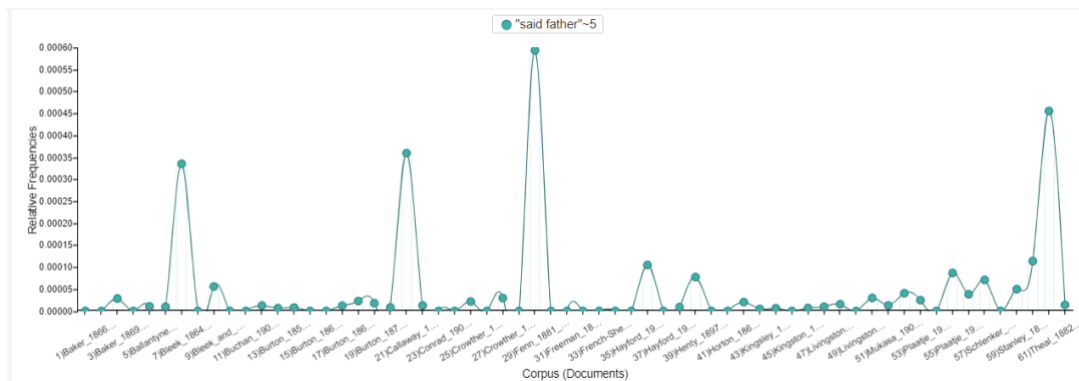


Fig. 13. Frequency graph of how often “said” and “father” appear within close proximity to each other throughout the corpus (file name: omvData_0029)

Searching through other words that had been identified as words that could indicate who is speaking throughout these texts, many of the most commonly collocated terms had similar results to those of “said” – they largely indicated that the texts of the corpus were dominated by male voices, or at least, male characters were speaking.

However, there were instances when using this tool that indicated that certain words, though commonly used as a speaking verb, were not used in this way throughout the corpus – even after searching through some of these words individually earlier in the research process. For example, while “called” was a commonly occurring word throughout the corpus, it was actually more likely to be used to describe “discoveries” that were made and things that people were learning while they visited Africa, names of communities that they talked to, or telling the name of a specific person that they encountered.

Documents Phrases Collocates			
Term	Collocate	Count (context)	
<input checked="" type="checkbox"/> called	people	78	
<input type="checkbox"/> called	little	70	
<input type="checkbox"/> called	large	67	
<input type="checkbox"/> called	men	65	
<input type="checkbox"/> called	country	65	
<input type="checkbox"/> called	small	64	
<input type="checkbox"/> called	river	62	
<input type="checkbox"/> called	arabs	61	
<input type="checkbox"/> called	said	60	
<input type="checkbox"/> called	man	58	
<input type="checkbox"/> called	came	57	
<input type="checkbox"/> called	mr	53	
<input type="checkbox"/> called	king	53	
<input type="checkbox"/> called	old	51	
<input type="checkbox"/> called	white	49	
<input type="checkbox"/> called	great	48	
<input type="checkbox"/> called	near	47	
<input type="checkbox"/> called	west	44	
<input type="checkbox"/> called	went	44	
<input type="checkbox"/> called	village	42	
<input type="checkbox"/> called	chief	42	

Fig. 14. List of words that appear most frequently within 5 words of “called” (file name: omvData_0031)

Ultimately, because many of the majority of the results for “called” were not about who was speaking, they could not be included in the analysis of the people or character who are able to speak throughout the corpus.

C. Summary

The initial analysis in Voyant researched words that are, currently, commonly associated with colonialism and imperialism, including: “take,” “took,” “wild,” and “old,” as well as religious terms like “god,” “christ,” and “religion.” The results of searching “take” and “took” in the corpus suggested that these words tended to correlate with physical bodies and life, including men, dead, and children. Results also suggested that when “wild” was used in the corpus, it was often accompanied by negatively-connotated words, like “brutal.” Religious terms did not appear to be discussed as often as expected throughout the corpus.

Voyant’s “Topic Modeling” tool uncovered topics or themes throughout the corpus that may not have been obvious upon reading all of the documents. Many of these clusters suggest the ways in which Africa and the expeditions by the British were discussed throughout the texts.

The results from research into who is speaking in the *One More Voice* produced thus far suggest that primarily British male characters or British male people’s voice dominated the narratives included in the corpus. Some of the results did indicate, however, that other voices were included in certain genres. For example, mothers and children, and even fathers, were more likely to speak in works of folklore or nursery rhymes.

V. Learning Python

A. Results

While Voyant is incredibly user-friendly, well-designed, and contains several different tools that can be used to analyze a corpus, a more nuanced tool or programming language such as [Python](#) may be necessary for a deeper analysis and more specific research questions. Python is a commonly used programming language for text analysis and natural language processing that uses a series of commands and functions to output data. For this project, several resources were used to learn the fundamentals of Python and of natural language processing using Python, including *Bite-Sized Python* by April Speight⁴, *Coding for Kids: Python* by Adrienne B. Tacke,⁵

⁴ April Speight, *Bite-Size Python* (Indianapolis: John Wiley & Sons, Inc., 2020).

⁵ Adrienne B. Tacke, *Coding for Kids: Python* (Emeryville: Rockridge Press, 2019).

and several lessons and courses on [CodeAcademy](#).⁶ Using these resources, I was able to understand how Python both alone and via use of the [Natural Language Toolkit](#) – a set of natural language processing tools that can be imported into Python – can be used to analyze a corpus, *and* how *One More Voice*'s own corpus may benefit from being imported into and analyzed in Python.

Through these resources and particularly through practicing with Python natural language processing coding through CodeAcademy, I learned about the following tools. It is crucial to keep in mind that these tools are not perfect and that the usefulness of each largely depends on the corpus, the research questions being asked about the corpus, and the goals of analysis. Several of the following tools can be found in the Natural Language Toolkit for Python:

- **Noise Removal** – clears the document of any formatting.
- **Tokenization** – breaks text down into smaller units and includes:
 - **Stemming** – removes prefixes and suffixes from words in the corpus.
 - **Lemmatization** – breaks down words to their roots.
- **Named Entity Recognition** – identifies proper nouns in a text or texts.
- **Part of Speech Tagging** – identifies parts of speech in a text or texts.
- **Dependency Grammar Trees** – visualizes the relationship of words within a sentence.
- **Language Models** – set of trained computer programs that calculate the probability that a word or phrase will be used. Language models must first be trained on a corpus or text in order to make predictions. Example language models include:
 - **Unigram Model/Bag of Words** – counts the number of times each word appears in a text or corpus.
 - **N-gram Model** – searches for a series of n units (in the case of text analysis, a phrase made up of a specified number, n , words) and predicts the likelihood that that string of units would appear within the entirety (in the case of text analysis, the probability that a phrase or string of words would appear within that text or corpus).
 - **Bigram Model** – searches for the most likely occurring two-word phrases in the corpus; type of language model, where n in an n -gram model is equal to two.
- **Topic Modeling** – finds topics within a text or corpus that may not otherwise be obvious and includes:
 - **Latent Dirichlet Allocation (LDA)** – finds topics within a text by generating a set of words that frequently appear in proximity to each other throughout a text or corpus

⁶ Code Academy, “Apply Natural Language Processing with Python.,” n.d., <https://www.codecademy.com/learn/paths/natural-language-processing>. and Code Academy, “Learn Python,” n.d., <https://www.codecademy.com/learn/learn-python-3>.

B. Process

Before learning about coding the basics of natural language processing, I used the books by Tacke and Speight (see prior subsection) to gain a foundational knowledge of Python. Through these books, I learned the basic terms of Python coding – functions, strings, commands, variables, etc. – as well as the structure for basic coding. While the best understanding of any coding or markup language comes with time, troubleshooting, and experience working on multiple projects, a foundational understanding of Python is needed to understand *how* the text processing tools in Python may work, especially once a corpus has been inputted.

In this regard, CodeAcademy was particularly helpful resources for further understanding Python, for learning about natural language processing and coding, and for getting to see how the tools actually function on sample text. While the tools were not specifically tailored to *One More Voice*'s own computational text analysis project, of course, there were tools that, with more knowledge and practice, could be further adapted in Python to fit the specific analysis that could be performed on the corpus.

CodeAcademy lessons are particularly useful because the user is introduced to the code, has a chance to examine the practice code, then is asked to run the practice code to understand how it works on a sample text. The lesson then guides the user through adjusting the code to achieve a certain objective. Users cannot move on to the next portion of the course until they have mastered the current portion and have generated the correct code. For the Natural Language Processing course, CodeAcademy has already important tools from the Natural Language Processing Toolkit and so teaches users how to use tools that have already been created, tested, and revised instead of teaching the users how to code similar tools from scratch. Finally, after completing the lessons, users are provided with an article about current critical discussions surrounding the information just learned. In the case of the particular lesson I used, the recommended reading was the book *Algorithms of Oppression* by Safiya Umoja Noble. After the recommended reading, users are given a quiz to complete the course, to review the lesson, and test the knowledge just developed.

C. Summary

Through research in Python workbooks and CodeAcademy courses, I was able to learn the basics of Python and of performing natural language processing in Python. Because of this, I was better able to understand the possibilities of what can be analyzed within the *One More Voice* corpus and how to go about doing that, but also to understand better the processes and functions of the Voyant tools used to analyze the corpus because of the learning about the coding for these tools while using CodeAcademy

For those who have experience using Python, using the Natural Language Toolkit along with Voyant may provide a more well-rounded analysis because tools can be customized to a specific research question. All researchers, however, should be aware that some of Voyant's tools are still experimental and likely need more development, according to Voyant's [tool index](#). That being said, using both Voyant and Python could allow for more thorough research to be completed because the corpus has been run through two similar tools.

VI. Conclusion

By consulting the *One More Voice* bibliography and literary scholars in Victorian-era literature, a corpus was created with the intention of being analyzed through computational text analysis by the *One More Voice* contributors and also made available through GitHub so that other scholars would be able to utilize the corpus for their research too. The corpus contains first editions of the texts, and, in rare cases where the first edition could not be found, the earliest edition that could be found was included in the corpus. The metadata for each work was collected in Zotero and a spreadsheet; the spreadsheet is included with the corpus.

Computational text analysis was then performed on the corpus using Voyant. Based on *One More Voice*'s own research interests and on the most frequent words of the corpus, I explored several topics, including words that may have indicated how themes of colonization and imperialism were embedded in the works, the results of Voyant's "Topic Modeling" tool after 2000 iterations were performed on the corpus, and who was able to speak throughout the words of the corpus.

Finally, to supplement my research in Voyant, I learned the basics of Python and natural language processing using Python to understand how a programming language may be used to perform more advanced computational text analysis.