Anne Salminen 014312564

Introduction to Methods in Digital Humanities

Autumn 2017

1. Looking for Words – Introduction to the Study

It is a truth universally acknowledged that a student of Comparative Literature in possession of an adaptation must be in want of a sound method for comparing the adaptation with the original text.

In this particular study, all of the texts that form the dataset are adaptations as they are translated in English and Finnish. It is worthwhile, however, to compare the texts and inspect the differences as well as similarities between them; the texts are slightly different versions of the original one and translated by different translators. Computational methods offer fine tools for this kind of distant-reading.

I analysed word frequencies in four versions of *The Arabian Nights*. The English works are *A plain and literal translation of the Arabian nights entertainments, now entituled The Book of the Thousand Nights and a Night*, Volume 1 (2016, transl. Richard F. Burton) and *Fairy Tales From The Arabian Nights* (2005, transl. E. Dixon). The Finnish works are *Tuhannen ja yhden yön tarinoita Suomen lapsille II* (2008, transl. Helmi Krohn) and *Tuhat ja yksi yötä* (2015, transl. Valfrid Hedman). All of the texts were derived from Project Gutenberg.

Although it was not clear from the beginning, the most important tools utilized were R, Word and Voyage. It might not be the conventional way of reporting to include the failings and somewhat trivial details of the process in the report. However, that is exactly what I will do. The decision is based on a number of reasons. First of all, as I am eager to learn more coding and consider this project as one valuable step on the slowly-revealing path, I believe it is most useful to reflect on the process. Secondly, my failings are one kind of metadata that might prove useful for the teacher (or any reader) in the future. Thirdly, I want to provide a view on the effort I put into this project.

2. Finding Paths, Losing Paths – Description of the Research Process

I have no previous coding experience but have been learning Python this autumn. Therefore, I initially worked on Python. The codes I tried to create and combine were ultimately inoperative but did manage to do some of the tasks I aimed to execute: for instance, I managed to make the text lower key and strip out punctuation characters. For some unfortunate reason I could not manage to strip out the digits nor could I manage to write the results in a csv-file (I managed to create a file but it remained empty no matter what primitive magic I tried). At this point I also had a different corpus: I wanted to compare different translations of the *Bible*. The ability to remove digits would have been crucial as all the verses are numbered. As I failed to do that, I had to give up *Bible*. I then chose to compare Finnish translations of the *Arabian Nights* that were mentioned earlier. I found the corpus intriguing as one of the texts was aimed for children and the other one for adults.

After some hours and a few invisible tears of frustration I chose to use R. I have been learning R with the help of Matthew L. Jockers's excellent book *Text Analysis with R for Students of Literature* (2014). With the help of the book and the internet I managed to create a code with which I could count word frequencies and remove Project Gutenberg related metadata and footnotes. R struggled with *vöcäbyläry* and, for instance, interpreted *ä* as one of the most frequent words. I tried to find help online but failed miserably

and initially gave up hope; that is when I looked up the English translations. However, it occurred to me later that I could use a simpler method. I copied the Finnish texts on Word and replaced letters “ä” with “a” and “ö” with “o”. After that each text was ready for R:

Code

```
text.v <- scan("lapsille_siivottu.txt", what="character", sep="\n")

start.v <- which(text.v == text.v[1]) #At first the value was different because I used the command to ignore
#Project Gutenberg headings but as I later removed them manually, I changed the value to 1.

end.v <- which(text.v == text.v[2794]) #See above.

start.metadata.v <- text.v[1:start.v -1]

end.metadata.v <- text.v[(end.v+1):length(text.v)]

metadata.v <- c(start.metadata.v, end.metadata.v)

novel.lines.v <- text.v[start.v:end.v]

novel.v <- paste(novel.lines.v, collapse=" ")

novel.lower.v <- tolower(novel.v)

text.words.l <- strsplit(novel.lower.v, "\\W")

text.word.v <- unlist(text.words.l)

not.blanks.v  <-  which(text.word.v!="")

text.word.v <-  text.word.v[not.blanks.v]

text.freqs.t <- table(text.word.v)

sorted.text.freqs.t <- sort(text.freqs.t , decreasing=TRUE)

sorted.text.freqs.t[1:100]
```

As the code implies, I figured out the 100 most frequent words in the text files, respectively. I downloaded tm-package and tried to utilize stop words but could not make it work. Therefore I was forced to apply a more clumsy method: I again copied the text on Word and utilized the Find and Replace -tool to replace all the “extra words” (and, or, etc) with blanks manually. It was not too laborious but in the future I would definitely rather use ready code, especially if I had a larger corpus.

After processing, re-processing, failing, re-reprocessing and so forth, the texts were ready for the final tool: Voyage. I uploaded each of the processed text files on the site and created word clouds. In the Finnish texts personal pronouns seemed to be overwhelming so in the future one might want to remove them from the dataset. I chose not to do it this time because I wanted to maintain the division into sexes in the English datasets.

3. Results

The word clouds displayed clear differences between the texts' most frequent words so I utilized Voyage's graph line tool and chose the same words for inspection in all of them: "king", "queen", "prince", "princess" and "allah". Interestingly, "allah" was not mentioned in *Fairy Tales From The Arabian Nights* nor *Tuhannen ja yhden yön tarinoita Suomen lapsille* which both seem to be aimed for children. In the latter one "prince" and "princess" correlated which implies that the characters have a social connection in the text. There is not as clear a connection between any of the characters in any other text. The graphs vary greatly when it comes to frequency of an invidual words at certain segments of the document. There are several reasons. First of all, the texts were modified during the translation and adaptation processes. Secondly, the datasets are of different lengths and contain a different amount of unique terms (information derived from Voyage):

A plain and literal translation of the Arabian nights entertainments, now entituled The Book of the Thousand Nights and a Night: 108,916 total words and 12,924 unique word forms.

Fairy Tales From The Arabian Nights: 58,932 total words and 5,243 unique word forms.

Tuhannen ja yhden yön tarinoita Suomen lapsille: 18,842 total words and 6,927 unique word forms.

Tuhat ja yksi yötä: 44,604 total words and 15,257 unique word forms.

It is worth noting that the texts for children contain considerably less unique word forms compared to the adult versions. Moreover, the Finnish adult version contains less total words and more unique words than its English equivalent. That might be caused by the language because Finnish is an agglutinative language and English is not.

4. Reproducability

I provide the links to Voyage here. The texts can still be further studied.

All of the original texts can be found and downloaded on Project Gutenberg. The datasets I ended up with are attached with the submission file.

A plain and literal translation of the Arabian nights entertainments, now entituled The Book of the Thousand Nights and a Night: https://voyant-tools.org/?corpus=ba426cd94b98ed62d75c19f2dc581d77

Fairy Tales From The Arabian Nights: https://voyant-tools.org/?corpus=4c4edda45635ad2e9ecc918a0b61e9f6

Tuhannen ja yhden yön tarinoita Suomen lapsille: https://voyant-tools.org/?corpus=51bdd5f03734b061ffbe8fd317d3ceac

Tuhat ja yksi yötä: https://voyant-tools.org/?corpus=658726f39d79a003246886aa3a4d05aa

5. In the future

This research could be taken forward by developing the coding part and adjusting the vocabulary. As I mentioned earlier, the personal pronouns are not necessarily the most important and interesting words. Furthermore, in this kind of research the original text(s) would bring invaluable insight into the subject. I would be most interested in gaps (like the ones we encountered with the word "Allah" in children's texts):

what has been left out of the adaptations, where and when. That might bring us one step closer to understanding why.