“A mere maiden”: Exploring Lúthien Tinúviel’s relationship with dance and song with word dispersions, correlations, topic modelling, and fuzzy matching

Christina Nguyen

INF2010: Natural Language Processing

Professor Rohan Alexander

April 24, 2022

Word count: 4638

Commented [CN1]: Pre-submission checklist:
- All "Luthien Tinuviel" is accented
- MLA formatted
- Images are all subtitled properly.
- Word count?
- Update page numbers in contents section
- Proper links to code, reading notes, explanation of how it all works
- Change all "we" to "I", or vice versa. Pick one to stick with.

- Change all italics of *singing* and *dancing* to quotation marks "singing" and "dancing", or vice versa. Pick one to stick with.
- Check that methodology summary matches what you actually ended up doing.

Commented [CN2R1]: Turn everything into present tense!

Contents

Abstract

The character of Lúthien Tinúviel has frequently been described as being central to the entire creation of the Eä universe. As previous scholars have identified, the two major sources of her power are dance and song, which are sometimes described as extensions of her femininity, though this is hotly debated and often outright rejected. Clare Moore, in her 2021 article *A song of greater power*, contends that J.R.R. Tolkien "increasingly […] establishes Lúthien as a figure of power" as the story is written and re-written, wielding these two art forms (song and dance) as expressions of self and influence (Moore). Tolkien also "increase[es] her agency and autonomy", and therefore "presents her as the foremost figure of his entire legendarium by establishing her influence over the history that comes after her," and part of that power comes from song and dance (Moore). Following on Moore's close reading work comparing the five major texts telling the story of Lúthien Tinúviel, I create a corpus for use in R; look at term frequency; create dispersion plots to visualize patterns of occurrences across the various versions of the story; and calculate correlation scores between *song* variants and *dance* variants. These are all ways to identify the true sources of her power and their evolutions over the five key manuscripts: *The tale of Tinúviel* (1917), *The Lay of Leithian* (1925), *Sketch of the mythology* (1926), *Quenta Noldorinwa* (1930), and *Quenta Silmarillion* (published 1977). The dispersion plots particularly probe Moore's argument that the art form of dance gives way to song over time. Lastly, I also consider what weaknesses these R tools bring to studying the Lúthien story, and make suggestions for a text analysis library for Tolkien studies.

1. Preliminary Matters

1.1. Hypothesis and research questions

Clare Moore's close-reading interpretation of Lúthien Tinúviel follows the evolution of the myth over five texts written and re-written over time. She focuses on the sources of Lúthien's power, i.e. song and dance, as an explanation for Lúthien's growing autonomy and power. In terms of outcomes, Moore found that earlier drafts showed Lúthien as less autonomous and less powerful. Subsequent revisions and drafts showed the evolution of the character into a powerful, active, and independent character who is central to the legendarium.

In order to make Moore's hypothesis machine testable (or, as it can be called, distant-reading and mixed-method reading), I focus specifically on one section of her arguments. Moore writes about a particular scene wherein "J.R.R. Tolkien's shift from dance [as the main source of Lúthien's power] to song in this scene reveals a focal shift […] Vink […] noted that the ratio between song-related words shifts dramatically between versions, with a 4:1 song:dance ratio in the *Lay* compared to a 16:1 ratio in chapter nineteen of *The Silmarillion*" (Moore). Indeed this brings up the possibility of expanding this type of "word count" (laterally called *token counts* or *term frequency* in natural language processing) test to study, for *each* of the five texts:

- the term frequency of "danc*" (which includes "dancing," "dances," "danced," etc.). Hereafter I will refer to these as *dance* variants.
- the term frequency of "sing*" and "song*" (which includes "singing", "songs," etc.) Hereafter I will refer to these as *sing* variants.

Note that above, every asterisk mark indicates a *wildcard function*, which in natural language processing (henceforth referred to as NLP) indicates unknown characters in a text value; this is useful for locating multiple items with similar, but not identical data. In this first case, we see a search for all the possible conjugations of the verb *dance.*

Future research could continue to probe Moore's arguments about similar passages across the five texts using fuzzy matching and n-grams.

1.2.The Lúthien Tinúviel story and myths

Why look at five retellings of the Lúthien Tinúviel story? Why identify the similarities and differences; why does this change the way we see the story? Why call it a myth? Myths are living, evolving, and growing narratives. We cannot expect myths to remain the same generation after generation – many attributes change, sometimes for reasons that we cannot understand in hindsight. To demand that a myth, particularly those *deliberately designed* as myths (as the case is here), be faithful to some "original" is dangerous and detrimental to the retelling performance, and also brings serious implications to manuscript studies as a whole. Thus we reject what is known as the *fidelity discourse*; we reject the idea that there is an original version to which we must adhere (Bortolotti and Hutcheon 445). A myth is about performance, about being spoken, told, about being received by an audience, and by changing. Crucially, as William Uricchio writes, "there is much more to be gained […] by exploring textual multiplicity, modification, and […] textual hacking as generative and interactive practices in their own right rather than as […] corruptions of an idealized end-state" (Uricchio from Pesce and Noto 155). Certainly, Tolkien understood this beyond the scope of the Lúthien Tinúviel story. In his own writing, there is the metatext of The Red Book, in which different characters are said to have recorded their adventures (from Bilbo in *The Hobbit*, it was then passed onto Frodo in *The Lord of the Rings* and finally to Samwise Gamgee). In fact, it is hinted that the story of *The Hobbit* itself is the first draft the Red Book (Ferré). In Tolkien's own studies, we know that he studied Old English narratives that were constantly retold and/or rewritten. So, there is no surprise in saying that understanding variation between myths (often by means of a variorum, like we can see by "Bilbo's footnotes" in *The Lay of Leithian*), and the context behind these changes, is essential to understanding Tolkien's obsession with myth and myth telling.

Again, of particular focus to this paper is the myth of Lúthien Tinúviel, the elven maiden whose story shaped those of her descendants' so powerfully. It is in the telling and retelling of her story that we witness the evolution of her character, both in qualities integral to her femininity and beyond. As Clare Moore notes, these qualities that make Lúthien "an active character and central to the legendarium [through] agency and autonomy" specifically are song and dance (Moore 7). To track changes across five moderately similar retellings of the same story, it is only sensible to use computational methods: it is in this manner that we discover minute similarities – not just in similar phrases, but in similar plot tensions, sentiments, and word frequency. Without computational methods, we are limited to the boundaries of human memory and pattern recognition; thus, given the tools built into R and various NLP packages, this research is more important, timely, and possible than ever.

1.3. The current state of related research

1.3.1. Measures of intertextuality

Despite widespread popularity in other sections of literature studies, very few scholars have used computational methods to study anything of Tolkien at all. What little there is, is not published in peer-reviewed journals. Techniques in this paper are inspired by Shmidman et al.'s paper *Identification of parallel passages across a large Hebrew/Aramaic corpus*, particularly when considering how fuzzy matching can help future research to identify longer phrases of parallels. James Tauber has created an exciting new project (which unfortunately is not yet published as an article) at digitaltolkien.com, wherein basic linguistics tests are performed on corpuses and some texts are marked up with XML. These fledging steps are key to Tolkien studies becoming more NLP-friendly and computational-linguistics-friendly in the near future. For now, in this paper, I make do by creating my own corpus solely of Lúthien Tinúviel texts and share it in the hopes that other researchers will join me in creating more themed corpuses, allowing Tolkienists to share the same datasets quickly and to replicate work.

1.3.2. Qualitative intertextuality in the Lúthien Tinúviel story

The academic conversation about Lúthien's power is extensive. Renée Vink is of especial interest here, but Melanie Rawls, Katarzyna Wiktoria Klag, Jack M. Downs, Edith Crowe, and Verlyn Flieger have all published various outstanding essays on Lúthien's world-building powers (Rawls) (Klag) (Downs) (Crowe) (Flieger). Vink, for example, finds that over the years, Tolkien turned the story away from one dominated by dance "into one which music and song gradually began to take over until only one dancing scene remained [, one that was] most personal to Tolkien himself" (Vink 257). Vink hand-counts words and phrases (which means there was a bit of quantitative work) that she identifies as related to music in a wider sense. Some of the problems she encounters in analysis are eliminated in my analysis too: for example, she notes that the three texts she studies (*Of Beren and Lúthien*, *The Tale of Tinúviel*, and *The Lay of Leithian)* are of widely different lengths; she counts it in pages. Thus Vink does not calculate relative frequencies of these words and phrases to make comparison more equitable, and she acknowledges this in saying "the three texts[' differing lengths] needs to be taken into consideration when looking at the results" (260). Though my paper focuses only on word searches to corroborate some of Vink's hand-counts, in future, the Shmidman techniques mentioned in 1.3.1. could be used to extrapolate Vink's phrase searches.

2. Quantitative methodology and results

2.1.Creating the clean corpus

The corpus includes the five seminal texts that Moore originally chose as representative of the Lúthien Tinúviel story. Again, they are, in chronological progression:

- *The Tale of Tinúviel* (1917), the first chapter of *The Book of Lost Tales Volume 2*
- *The Lay of Leithian* (1925),

- *Sketch of the mythology* (1926), also known as *The Earliest 'Silmarillion'* and the second chapter of *The Shaping of Middle-Earth*
- *Quenta Noldorinwa*, the third chapter of *The Shaping of Middle-Earth* (1930), and
- *Quenta Silmarillion*, the third part of *The Silmarillion* text (published 1977)

While there are at least nine different drafts of this story, these five are the ones that Christopher Tolkien uses to "compile the single volume *Beren and Lúthien*," the most whole telling of the story in a single manuscript (Moore).

For *The Tale of Tinúviel*, I have removed Christopher Tolkien's footnotes and have not included the second version (which is very close to the first) of the tale.

For *The Lay of Leithian*, I have removed Bilbo's forward and commentaries. Though this is a key contextual positioning of this version of the myth, indeed it is paratext (i.e. an element not part of the primary text), or what Gerard Genette calls *peritext*, rather than part of the myth itself (Genette). Its removal therefore will not affect the accuracy of the final corpus and textual analysis.

For *Sketch of the mythology*, I have removed the majority of Christopher Tolkien's commentaries, save those at the end of the chapter wherein he summarizes' Lúthien's plight. Like *Quenta Noldorinwa* (below), since it comes from *The Shaping of Middle-Earth*, the commentaries help provide context but are not central to the computational analysis. I have also only included sections 10 forward which have especial focus on Lúthien, rather than Morgoth and the evil surrounding Morgoth.

For *Quenta Noldorinwa*, I have removed Christopher Tolkien's footnotes and commentaries. Though the format of *Quenta Noldorinwa* is undoubtedly a compilation of nebulous sources, and the commentaries put context to these vignettes, again the commentaries are not part of the primary text and are thus erased from the final corpus. Additionally, only sections 10 and 11 of the *Quenta Noldorinwa* is included in the corpus, as sections outside of these do not cover the story of Lúthien Tinúviel in detail, and thus additional data would affect the accuracy of the final corpus and textual analysis.

For *Quenta Silmarillion*, I have included only chapter 19 of *The Silmarillion*, titled "Of Beren and Lúthien," obviously since this is the only chapter that exclusively focuses on Lúthien's adventures.

The corpus is made of five .txt files, which are easily parsed by read-in functions in R and may manipulated with any number of NLP packages. These files can be found at the stable link: shorturl.at/esLWZ.

2.2. Word frequencies

2.2.1. Absolute and relative frequencies of all of sing *and* dance *variants*

Shown below are the absolute and relative frequencies of each umbrella of variants. *Sing* variants include *sing, sang, singing, sings, sung,* and *song. Dance* variants include *dance, dancing, danced,* and *dances.*

Text	Absolute and relative frequency of *sing* variants	Absolute frequency of *dance* variants
The Tale of Tinúviel	24, 0.0018	30, 0.0023
The Lay of Leithian	80, 0.0037	22, 0.0010
Sketch of the mythology	4, 0.000018	1, 4.6e-05
Quenta Noldorinwa	9, 0.000018	1, 0.00032
Quenta Silmarillion	36, 0.0007	3, 8.1e-05

Figure 1. Absolute and relative frequencies of 'sing' variants and 'dance' variants across 5 texts.

As we can see, the extremely low amount of occurrences in some these instances means the values are almost trivial. This will affect how we perform the final mixed-method analysis.

```
#How many words are "dance"
dance_hits_v <- length(loweredtext[which(loweredtext=="dance")])
dance_hits_v #9 hits
#How many words are "dancing"
dancing_hits_v <- length(loweredtext[which(loweredtext=="dancing")])
dancing_hits_v #8 hits
#How many words are "danced"
danced_hits_v <- length(loweredtext[which(loweredtext=="danced")])
danced_hits_v #11
#How many words are "dances"
dances_hits_v <- length(loweredtext[which(loweredtext=="dances")])
dances_hits_v #2

#Therefore total number of all iterations of danc- is 9+8+11+2 = 30
totaldancehits_v = 30
total_words_v <- length(loweredtext)
totaldancehits_v/total_words_v #which tells us that "danc-" makes up 0.0023 of the whole Tale of Tinuviel

#How many words are "song"
song_hits_v <- length(loweredtext[which(loweredtext=="song")])
song_hits_v #12
#How many words are "sing"
sing_hits_v <- length(loweredtext[which(loweredtext=="sing")])
sing_hits_v #1
#How many words are "singing"
singing_hits_v <- length(loweredtext[which(loweredtext=="singing")])
singing_hits_v #2
#How many words are "sang"
sang_hits_v <- length(loweredtext[which(loweredtext=="sang")])
sang_hits_v#7
#How many words are "sings"
sings_hits_v <- length(loweredtext[which(loweredtext=="sings")])
sings_hits_v #1
#How many words are "sung"
sung_hits_v <- length(loweredtext[which(loweredtext=="sung")])
sung_hits_v #1

#Therefore total number of all iterations of sing- is 12+1+2+7+1+1 = 24
totalsinghits_v = 24
totalsinghits_v/total_words_v #which tells us that "sing-"'s iterations makes up 0.0018 of the whole Tale of Tinuviel
```

Figure 2. A sample calculation from R of the absolute and relative frequencies for both the *sing* variants and *dance* variants, using only "The Tale of Tinúviel." For the chart above with all five texts, this calculation was simply repeated for the remaining four texts.

2.2.2. Word dispersions across the story

While the section above is good at telling us *how much* of the story *sing* variants and *dance* variants makes up, it does not tell us much about *where* they occur in the story. For this test I will

treat the order in which the words appear in the text as a measure of time, also called *novelistic time*. That means that the first word in every text is the index is n = 1.

```
#create novelistic time index

n_time_v <- seq(from = 1, to = length(loweredtext))

#1b1 identify at which index points the word "dance" occurs
dance_v <- which(loweredtext == "dance")
dance_count_v <- rep(NA, times = length(n_time_v))
dance_count_v[dance_v] <- 1

plot(dance_count_v, main = "Dispersion plot of 'dance' in TOT",
     xlab = "novelistic time", ylab = "dance", type = "h", ylim = c(0,1), yaxt='n')

#1b2 identify at which index points the word "dancing" occurs
dancing_v <- which(loweredtext == "dancing")
dancing_count_v <- rep(NA, times = length(n_time_v))
dancing_count_v[dancing_v] <- 1

plot(dancing_count_v, main = "Dispersion plot of 'dancing' in TOT",
     xlab = "novelistic time", ylab = "dancing", type = "h", ylim = c(0,1), yaxt = 'n')

#1b3 identify at which index points the word "danced" occurs
danced_v <- which(loweredtext == "danced")
danced_count_v <- rep(NA, times = length(n_time_v))
danced_count_v[danced_v] <- 1

plot(danced_count_v, main = "Dispersion plot of 'danced' in TOT",
     xlab = "novelistic time", ylab = "danced", type = "h", ylim = c(0,1), yaxt = 'n')

#1b4 identify at which index points the word "dances" occurs
dances_v <- which(loweredtext == "dances")
dances_count_v <- rep(NA, times = length(n_time_v))
dances_count_v[dances_v] <- 1

plot(dances_count_v, main = "Dispersion plot of 'dances' in TOT",
     xlab = "novelistic time", ylab = "dances", type = "h", ylim = c(0,1), yaxt='n')

#overlay/combine all the above dance variants together
```

Figure 3. A sample dispersion plot calculation from R for all *dance* variants, using only "The Tale of Tinúviel." For the chart below with all five texts, this calculation was simply repeated for the remaining four texts and for *song* variants as well.

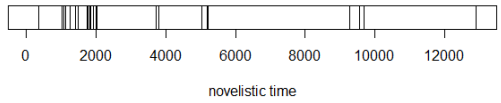
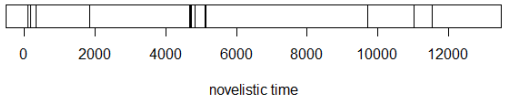
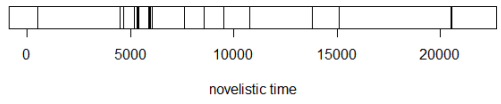
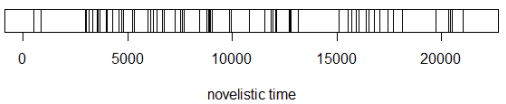
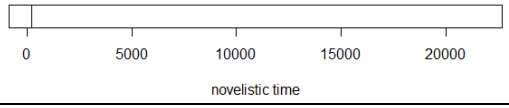
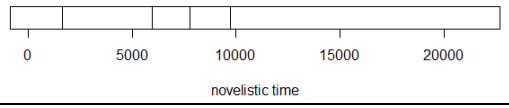
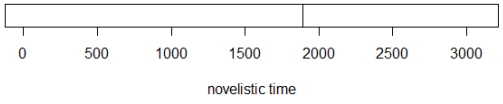
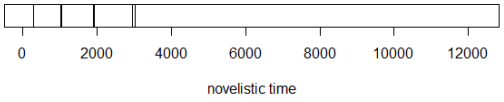
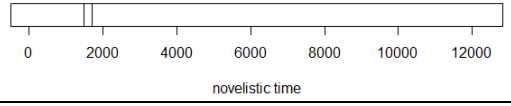
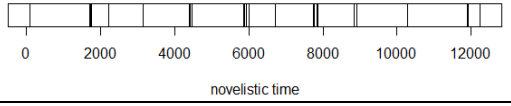
Text	Dispersion plot for *dance* and *dancing* and *danced* and *dances*	Dispersion plot for *song* and *sing* and *singing* and *sang* and *sings* and *sung*
The Tale of Tinúviel		
The Lay of Leithian		
Sketch of the mythology		
Quenta Noldorinwa		
Quenta Silmarillion		

Figure 4. Dispersion plots for 'dance' and 'song' variants across the 5 texts.

2.2.3. *Correlation across the stories*

While the section above is good at telling us where *sing* variants and *dance* variants occur in the story, it is dangerous to assume that we can already make a statement about *correlation* between occurrences of them. Instead, using the frequency data we have compiled for these two variants, we run correlation tests to see if there is a *statistically significant relationship* between them. A correlation analysis is designed to determine the extent to which there is a linear dependence between two variables, i.e. the relationship between occurrences of *sing* variants and *dance* variants.

To simplify, the question here is: "To what extent does the usage of *sing* variants change in relation to the usage of *dance* variants, and vice versa?"

2.2.3.1. Correlation for *The Tale of Tinúviel*

R provides a simple function, `cor`, for finding this correlation value. First, we break each of the texts into equally-sized chunks of 200 words (herein each chunk will be referred to as *chunk*), with the function `chunk_text()`, and store it in a variable called `chunky_text`.

```
#Split into chunks of 200 words each
library(tokenizers)
library(tidytext)
library(readtext)
library(tibble)
text_v <- readtext("C://Users//chris//Downloads//TaleofTinuviel.txt")
text_v
text_v <- tolower(text_v)
text_v

chunky_text <- text_v %>%
  chunk_text(200) #chunked sucessfully!
```

Figure 5. Splitting a text into chunks of 200 words each.

Then, we use `str_count()` to find how many times *song/sang/sing/sings/sung* occurs in each chunk, and likewise with the variants of *dance.*

```
> str_count(chunky_text, pattern = "song")
 [1] 0 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 1 0 0 1 0 0 0 0 0 2 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
> str_count(chunky_text, pattern = "sing")
 [1] 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0
[50] 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0
> str_count(chunky_text, pattern = "sang")
 [1] 0 2 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> str_count(chunky_text, pattern = "sung")
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> str_count(chunky_text, pattern = "singing")
 [1] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> str_count(chunky_text, pattern = "sings")
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

> str_count(chunky_text, pattern = "dance")
 [1] 0 1 0 0 0 1 4 1 0 2 1 7 2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 1 0 0 0 0 0 1 0 7 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0
> str_count(chunky_text, pattern = "dances")
 [1] 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
> str_count(chunky_text, pattern = "dancing")
 [1] 0 0 1 0 0 0 0 0 2 0 2 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
> str_count(chunky_text, pattern = "danced")
 [1] 0 1 0 0 0 0 2 1 0 0 1 2 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 6. Integers for all the variants across the 83 chunks.

To simplify this, we add the variables of *song*, *sing*, *sang*, and *sung* into a single variable called song_variants_count; and we add the variables of *dance*, *dances*, *dancing*, and *danced* into a single variable called dance_variants_count.

```
> song_variants_count <- song_temp+sing_temp+sang_temp+sung_temp+singing_temp+sings_temp
> song_variants_count
 [1] 0 4 2 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 8 3 2 0 0 1 0 0 0 0 1 0 0 0 0 2 1 0 1 0 0
[50] 0 0 0 1 0 0 2 0 1 1 0 0 4 0 0 1 0 0 0 2 0 0 0 2 0 0 0 0 0 0 0 1 0 0

> dance_variants_count <- dance_temp+dances_temp+dancing_temp+danced_temp
> dance_variants_count
 [1] 0 2 1 0 0 1 7 2 2 2 4 9 3 0 1 0 0 0 0 0 0 0 3 0 1 0 0 0 0 0 2 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[50] 0 0 0 1 0 0 0 0 0 1 0 9 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0
```

Figure 7. Combining all the variant terms of each category ('song' and 'dance') into its umbrella.

We will also combine these two simplified integers into a single matrix of 83 rows and 3 columns (those columns are chunk number, number of song variants, and number of dance variants). This will be called bound_both_variants.

```
> head(bound_both_variants)
  song_variants_count dance_variants_count
1                   0                    0
2                   4                    2
3                   2                    1
4                   1                    0
5                   0                    0
6                   1                    1
```

Figure 8. Binding song_variants_count and dance_variants_count into a single matrix.

Without the chunk number column, it will be a matrix of two columns, so the result of calling cor is a new matrix containing two rows and two columns. The values in those cells are the correlation values:

```
> cor(bound_both_variants)
                     song_variants_count dance_variants_count
song_variants_count           1.00000000           0.04237823
dance_variants_count          0.04237823           1.00000000
```

Figure 9. Correlation values for bound_both_variants.

It is no surprise to see that song_variants_count (on the x) is perfectly correlated, at a value of 1.00, with song_variants_count (on the y). It is also no surprise to see that dance_variants_count (on the x) is perfectly correlated, at a value of 1.00, with dance_variants_count (on the y). The interesting detail is that dance_variants_count and song_variants_count's correlation is ~0.04. This Pearson Product-moment correlation coefficient tells us two things:

- There is a positive correlation, since our value is positive
- The positive correlation is weak, since the value is relatively close to 0 and not 1.

The weak correlation means, and *only* means, that as the usage of dance increases, song does not decrease significantly. This does not mean that one word is more or less important than the other or that we can read into textual sentiments yet – that remains to be seen with the mixed-reading (i.e. combining our results here with close reading). We always must contextualize the coefficient.

2.2.3.2. Correlation for *The Lay of Leithian*

We repeat the steps used in 2.2.3.1 to calculate correlation. Create the integer for each of the variants:

```
> song_variants_count
  [1] 0 0 2 0 1 1 0 0 0 0 0 0 0 0 0 0 1 2 3 1 2 4 2 1 2 1 2 1 0 3 2 1 0 0 1 1 1 0 1 2 1 0 0 0 1 0 1 4 2
 [49] 0 0 0 2 0 4 5 9 0 0 0 1 5 1 0 0 0 1 0 0 0 2 0 3 2 0 0 0 0 6 2 1 0 0 0 0 0 0 0 0 0 0 0 0 2 5 3 1
 [97] 0 0 0 5 0 0 3 2 2 0 0 2 3 2 2 1 0 1 1 3 0 0 2 0 2 0 2 0 1 0 1 1 2 0 1 0 0 2 0 0 0 3 2 2 0 0 3 0
[145] 0 0 0
> dance_variants_count
  [1] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 1 0 3 5 0 0 7 2 0 0 0 0 0 0 0 0 0 0 1 0
 [49] 0 0 0 0 1 0 0 0 0 0 3 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 2 0 0
 [97] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0
[145] 0 0 0
```

Figure 10. Integers for all the variants across the 147 chunks.

Bind the variants and convert the variable into a matrix where [,1] is song_variants_count and [,2] is dance_variants_count:

```
> head(bound_both_variants)
     [,1] [,2]
[1,]    0    0
[2,]    0    0
[3,]    2    1
[4,]    0    0
[5,]    1    0
[6,]    1    0
```

Figure 11. Binding song_variants_count and dance_variants_count into a single matrix.

Finally, perform the cor():

```
> #Correlation test
> cor(bound_both_variants)
           [,1]       [,2]
[1,] 1.00000000 0.05150891
[2,] 0.05150891 1.00000000
```

Figure 12. Correlation values for both variants.

Again, we see a weak positive correlation between the *song* and *dance* variants.

2.2.3.3. Correlation for *Sketch of the mythology*

We repeat the steps used in 2.2.3.1 to calculate correlation. Create the integer for each of the variants:

```
> song_variants_count
  [1] 0 1 1 1 0 0 0 0 2 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 2 0 0 0 0 0 1 0 0
 [49] 0 1 1 0 0 1 0 1 0 0 1 0 3 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [97] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

> dance_variants_count
  [1] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [49] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [97] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 13. Integers for all the variants across the 119 chunks.

Bind the variants and convert the variable into a matrix where [,1] is song_variants_count and [,2] is dance_variants_count:

```
> head(bound_both_variants)
     [,1] [,2]
[1,]    0    1
[2,]    1    0
[3,]    1    0
[4,]    1    0
[5,]    0    0
[6,]    0    0
```

Figure 14. Binding song_variants_count and dance_variants_count into a single matrix.

Finally, perform the cor():

```
             [,1]        [,2]
[1,]  1.00000000 -0.04817629
[2,] -0.04817629  1.00000000
```

Figure 15. Correlation values for both variants.

2.2.3.4. Correlation for *Quenta Noldorinwa*

We repeat the steps used in 2.2.3.1 to calculate correlation. Create the integer for each of the variants:

```
> song_variants_count
 [1] 0 4 0 0 0 2 0 0 0 3 0 1 0 0 2 1

> dance_variants_count
 [1] 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0
```

Figure 16. Integers for all the variants across the 16 chunks.

Bind the variants and convert the variable into a matrix where [,1] is song_variants_count and [,2] is dance_variants_count:

```
> head(bound_both_variants)
     [,1] [,2]
[1,]    0    0
[2,]    4    0
[3,]    0    0
[4,]    0    0
[5,]    0    0
[6,]    2    0
```

Figure 17. Binding song_variants_count and dance_variants_count into a single matrix.

Finally, perform the cor():

```
> cor(bound_both_variants)
          [,1]      [,2]
[1,] 1.0000000 0.4570188
[2,] 0.4570188 1.0000000
```

Figure 18. Correlation values for both variants.

Here is a relatively stronger correlation compared to the previous correlation tests, at nearly positive 0.5.

2.2.3.5. Correlation for *Quenta Silmarillion*

We repeat the steps used in 2.2.3.1 to calculate correlation. Create the integer for each of the variants:

```
> song_variants_count
  [1] 2 0 0 0 0 0 1 0 4 0 0 1 0 0 0 1 1 1 0 0 0 0 7 0 0 0 0 0 0 6 3 0 0 1 1 0 1 1 0 5 0 0 0 0 1 2 0 0 0
 [50] 0 2 0 2 1 0 0 0 0 0 0 5 1 0
> dance_variants_count
  [1] 0 0 0 0 0 0 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [50] 0 0 0 1 0 0 0 0 0 0 0 0 0 0
```

Figure 19. Integers for all the variants across the 119 chunks.

```
     [,1] [,2]
[1,]    2    0
[2,]    0    0
[3,]    0    0
[4,]    0    0
[5,]    0    0
[6,]    0    0
```

Figure 20. Binding song_variants_count and dance_variants_count into a single matrix.

Finally, perform the cor():

```
> cor(bound_both_variants)
          [,1]      [,2]
[1,] 1.0000000 0.1688564
[2,] 0.1688564 1.0000000
```

Figure 21. Correlation values for both variants.

2.6 Sources of error

There is the assumption that, in searching for *sing*-based and *dance*-based words, Lúthien is the only person who is tied to the word. Does Beren sing? Does Morgoth sing? Does Thingol dance? In my computations, I have assumed that every occurrence of these terms automatically means Lúthien is the actor. In my close reading, I have not seen evidence to suggest otherwise, but humans are prone to error when hand-counting. Thus future research should create a way to computationally check if other characters are using these words, perhaps through a mix of topic models and cluster models.

Additionally, there are some cases where other words are mixed with *sing* and *dance* variants to negate them. For example, in *The Lay of Leithian*, section VI, "Daeron's flute/ and Lúthien's singing both were mute" (J.R.R. Tolkien, *Lay*, 61). Future research should create a way to computationally check if other terms, like "not," "mute," "silenced," etc. have an effect on the qualities of *sing* and *dance*.

Furthermore, the five texts have different formats and perhaps it is not fair to compare them so blandly. *Quenta Noldorinwa* is written in a vignette-style and is considerably more choppy than the other four texts; *The Lay of Leithian* is a poem and not prose. Future research should consider the formats and changes in terminologies used within each text to better assess correlation.

In future, a special corpus of texts (as a library within R) with all of Tolkien's short and long fictional stories would be ideal for Tolkienists. They should be categorized with all relevant metadata (date of publication, variation details, etc.) and support special characters. In the case of this paper, that corpus could include options to use either the short or culled version of each of the Lúthien Tinúviel texts (as described in my corpus creation step). In some cases keeping Bilbo's commentaries in the text could be useful for analysis. It would be time-saving to researchers to not manually remove the text as I had above in the cleaning process.

3. Mixed-method reading: matching the computations to the close reading

From the dispersion plots, we see three interesting trends:

- From *The Tale of Tinúviel*, there is significant clustering of *dance* variants at the beginning of the story. Yet there is no significant clustering for *dance* variants in any of the other four texts.
- From *The Lay of Leithian*, there is a significantly even spread of *song* variants across novelistic time. The same is true for *Quenta Silmarillion.* Yet for the other three texts, there are many less occurrences of these variants and they do not appear as well-spread. Even between *Lay* and *Silmarillion* there is a significant drop-off the number of variants' occurrences.
- *Sketch of the Mythology* and *Quenta Noldorinwa* has so few data points that it is difficult to make any conclusions about the relationship between *song* and *dance* (whether inverse or proportional) across novelistic time.

From these quantitative observations alone, it is difficult to say Tolkien placed greater emphasis on either *song* (or *sing*) and *dance* over time (i.e., with each rewriting) or even both. To the naked eye, *dance* has diminished over time to virtually nothing; and while *song* does appear regularly in the last text, it is not as frequent as in the earlier *Lay.* Additionally, though Vink found a 16:1 ratio of song:dance in *Quenta Silmarillion,* my calculations from Figure 1 shows that it is actually 36:3, i.e. 12:1. While, like Vink, this shows that song greatly outweighs dance in *Quenta Silmarillion,* this does add some precision and accuracy to the figure. Vink also noted that *Lay* had a 4:1 (= 4) song:dance ratio, and I found that to be 80:22, i.e. 40:11 (~ 3.63). Again our values are more or less approximate to each other, but my values add some precision and accuracy again. So Vink's, and Moore's, arguments about the shift in emphases from the early *Lay* to the late *Quenta Silmarillion* by extension are supported by statistics.

From the correlation tests, surprisingly we saw that *within* the texts (not between them, as we considered with dispersion plots), there was no strong correlation between the variants. *Dance* variants appeared in more or less the same places as *song* variants and neither was preferred (used enough) to outweigh the other significantly in the cor() tests.

4. Postludium

Thus, a computational reading has supported Moore's explanation of the evolution of Lúthien Tinúviel's characters over the five texts. Some weaknesses of computational reading have been identified for further consideration – indeed computational reading should not be discounted altogether for textual variants, but developed and improved. The reading shows that subsequent revisions and drafts did indeed show the evolution of the character into a powerful, active, and independent character who is central to the legendarium. Word count, dispersion plotting, and correlation tests are simply the baselines upon which subsequent tests can build on, with expanded word/theme choices and different hypotheses. All code can be found for remixing at shorturl.at/bsGKX. J.R.R. Tolkien himself noted that he was written enough work to give

scholars something to study for a generation or two. Here we are at the tail-end of that timeframe, and yet some see that there is much to be done still. The work I have presented here is hopefully a flame from which “a fire shall be woken” beyond even the center of the legendarium, the mere maiden Lúthien Tinúviel (J.R.R. Tolkien, *The Fellowship of the Ring*).

Works cited

Bartolotti, Gary R; Hutcheon, Linda, *On the origin of adaptations: rethinking fidelity discourse and "success": biologically* (The Johns Hopkins University Press , 2007).

Crowe, Eith, "Power in Arda: Sources, Uses, and Misuses," in *Perilous and Fair: Women in the Works and Life of J.R.R. Tolkien*, ed. by Janet Brennan Croft and Leslie A. Donovan.

Downs, Jack M., "Radiant and terrible: Tolkien's Heroic Women as Correctives to the Romance and Epic Traditions," in *A Quest of Her Own: Essays on the Female Hero in Modern Fantasy*, ed. By Lori M. Campbell (2014).

Ferré, V. (2021). The Red Book and Tolkien's "The Lord of the Rings": A Fantastic Uncertainty. *Mallorn*, *1*, 26–33.

Flieger, Verlyn, "The Music and the Task: Fate and Free Will in Middle-Earth," in *Green Suns and Faerie: Essays on J.R.R. Tolkien* (Kent, OH: Kent State University Press, 2012).

Genette, G. (1997). *Paratexts : thresholds of interpretation*. Cambridge University Press.

Interactivity and the Modalities of Textual-Hacking: From the Bible to Algorithmically Generated Stories. Routledge, 2016, pp. 165–79, https://doi.org/10.4324/9781315718330-aaaaa19.

Klag, Katarzyna Wiktoria, 'The Power of music in the tale of Beren and Lúthien by J.R.R. Tolkien,' in *Analyses, Rereadings, Theories*, ed. By Joanna Matyjazczyk and Maciej Wieczorek (2014).

Moore, C. (2021). A Song of Greater Power: Tolkien's Construction of Lúthien Tinúviel. *Mallorn: The Journal of the Tolkien Society*, *1*, 6–16.

Rawls, Melanie, 'The Feminine principle in Tolkien,' in *Perilous and Fair: Women in the Works and Life of J.R.R. Tolkien*, ed. by Janet Brennan Croft and Leslie A. Donovan.

Shmidman, A., Koppel, M., & Porat, E. (2016). *Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus*. https://doi.org/10.46298/jdmdh.1388

Tolkien, J.R.R. "The Lay of Leithian." n.d. *Thain's Book.* PDF. 24 April 2022. <https://thainsbook.files.wordpress.com/2015/07/the-lay-of-leithian.pdf>.

Tolkien, J. R. R. (John Ronald Reuel), and Christopher. Tolkien. *The Book of Lost Tales*. HarperCollins, 1994.

Tolkien, J. R. R. (John Ronald Reuel), and Christopher. Tolkien. *The Shaping of Middle-Earth : The Quenta, the Ambarkanta and the Annals ; Together with the Earliest "Silmarillion" and the First Map.* Allen & Unwin, 1986.

Tolkien, J.R.R. (John Ronald Reuel), and Christopher. Tolkien. *The Silmarillion.* Ballantine Books, 2002.

Tolkien, J.R.R. (John Ronald Reuel). *The Fellowship of the Ring*. Harper Collins, 2007.