“A mere maiden”: Exploring Lúthien Tinúviel’s relationship with dance and song with tf-idf scores and fuzzy matching

Christina Nguyen

INF2010: Natural Language Processing

Professor Rohan Alexander

March 25, 2022

Word count: 4000

Commented [CN1]: Pre-submission checklist:
- All "Luthien Tinuviel" is accented
- MLA formatted
- Images are all subtitled properly.
- 15-20 works referenced
- Word count: 4000 +- 10%
- 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]: Why doing these steps in this order will help us answer this question ()
What potential biases this has, what assumptions have I made,
What have I done to eliminate as much statistical bias as possible
What falls outside of this scope
How this feeds into previous researchers’ work – who has done what? Mini lit review

Commented [CN3]: https://www.youtube.com/watch?v=khsNsa3-VqQ&ab_channel=VanityFair for fuzzy matching

Contents

Commented [CN4]: Question for proofreaders: Should this section come before "hypothesis and research questions" or after?

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 Lúthien Tinúviel's 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; integrate the <dplyr>, <tidytext>, and <ggplot2> packages; calculate term frequency-inverse document frequency (tf-idf) scores 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;* and perform fuzzy matching. The tf-idf scores of the corpus tell us how important song- and dance- related terms are in each of the 5 texts as well as overall, which probes Moore's argument that the art form of dance gives way to song over time. The tf-idf scores also indicate that [XYZ], which partially rejects Tom Shippey's claim of [XYZ] in his [year] [article name]. The fuzzy matching supplements close reading attempts to identify similar passages between the manuscripts, especially in a systematic, descriptive and efficient manner that can only be pursued using computational methods. Using the program R to identify n-grams with varying lengths, and varying 'fuzzy factors', I expand on Moore's close reading where she identifies several key passages that have stayed consistent throughout the story. Naturally this fuzzy matching also contextualizes the tf-idf scoring, allowing me to tell a well-rounded narrative in the mixed-method reading. Lastly, I also consider what weaknesses these R packages bring to studying the Lúthien story, and make suggestions for a text analysis package designed for use on Tolkien texts, including a specially designed lexicon, designed to minimize those weaknesses.

Commented [CN5]: Did you use the stringr package? If so, mention that here.

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 5 texts written progressively forwards in 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 two sections of Moore's arguments. Firstly, 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* in natural language processing) test to study, for *each* of the five texts:

- the term frequency of "danc*" (which includes "dancing," "dances," "danced," etc.)
- the term frequency of "sing*" and "song*" (which includes "singing", "songs," etc.)
- calculating the inverse term frequency of the above terms

Note that above, every asterisk mark indicates a *wildcard function*, which in natural language processing 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.*

Secondly, Moore notes that "[t]he oscillation of who names Lúthien – herself or Morgoth – reveals that J.R.R. Tolkien's development of Lúthien's agency is not always a steady progression" (Moore). Expanding from this idea of oscillating word choices, I investigate if there are *other* direct passages of text that we can quote that show nearly exact same phrasing of words. Moore's close reading identified a few, but this was not the main focus of her argument – so it is up to us to do that. That can help us find the "essential differences" between the drafts' portrayal of the character, *aside* from song and dance alone. In essence this means using these machine-testing methods:

- Fuzzy matching using n-grams and regex, using several key scenes (climaxes) that occur repeatedly throughout the 5 texts, to identify what words changed

These scenes are:

- Beren (Lúthien's lover)'s first encounter with Lúthien
- Lúthien putting guards to sleep with song
- Lúthien's battle against the monster Morgoth
- The naming of Lúthien (herself or Morgoth)

Commented [CN6]: Not the right word.

The method here is similar to those described in Shmidman et al.'s paper identifying similar parallel passages across various versions of the Talmud (Shmidman et al.).

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, 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, and sometimes for reasons that we can. To demand that a myth, particularly those *deliberately designed* as myths, 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. A myth is about performance,

about being spoken, told, and 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 what […] textual hacking as generative and interactive practices in their own right rather than as […] corruptions of an idealized end-state" (Uricchio). Certainly, Tolkien understood this beyond the scope of the Lúthien Tinúviel story. In his own writing, we have the metatext of The Red Book, in which different characters have been 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), 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” (Moore) specifically are song and dance. To track changes across 5 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 fairly new R packages mentioned in the abstract, this research is more important, timely, and possible than ever.

1.3. The current state of related research

Vink
Schmidmin et al
Ferre
Tom Shippey on LT
James Tauber ngrams

2. Quantitative methodology and results

2.1.Creating the clean corpus

The corpus includes the 5 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

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 Tinuviel,* 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* was 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 5 .txt files, which are easily read by read-in functions in R and manipulated with any number of NLP packages. These files can be found at the stable link: shorturl.at/eltGU.

2.2. Word frequencies

2.2.1. Absolute and relative frequencies of all of "sing" and "dance"'s conjugations (iterations)

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

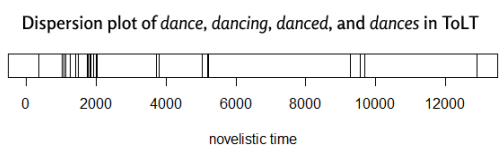
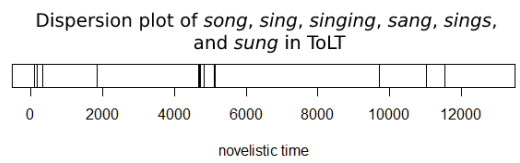
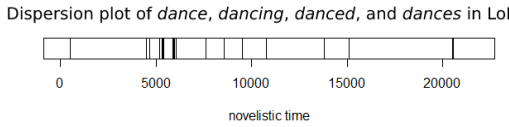
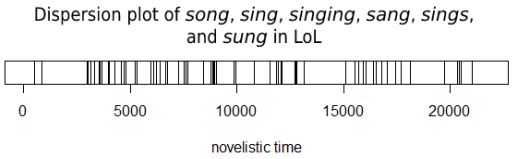
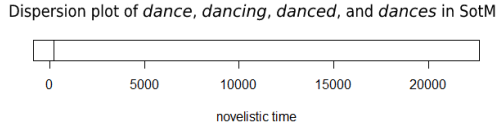
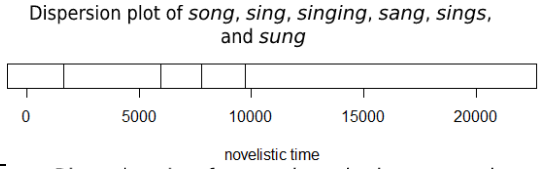
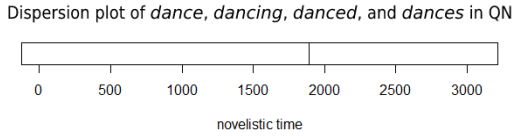
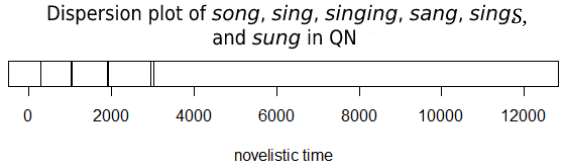
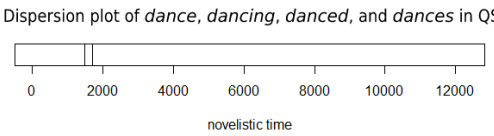
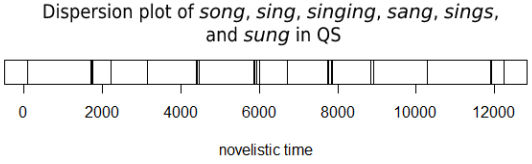
*As we can see, the extremely low amount of occurrences of these instances means the values are almost trivial. This will affect how we perform the final mixed-method analysis.

```
#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 1. A sample calculation from R of the absolute and relative frequencies for both the conjugations of "sing" and "dance", 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 up *how much* of the story "singing" and "dancing" 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, etc.

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*
Tale of Tinúviel		
Lay of Leithian		
Sketch of the mythology		
Quenta Noldorinwa		
Quenta Silmarillion		

2.2.3. *Correlation across the story*

While the section above is good at telling us *where* "singing" variants and "dancing" variants occur in the story, it is dangerous to assume that we can already make a statement about *correlation* between occurrences of "singing" variants and "dancing" variants. Instead, using the frequency data we have compiled for those two terms, we need to 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 "singing" variants and "dancing" variants.

To simplify, the question here is: "To what extent does the usage of *singing* variants change in relation to the usage of *dancing* variants?"

2.2.4. The Tale of Luthien Tinuviel

- Potential error: the term frequency of "danc*" (which includes "dancing," "dances," "danced," etc.)
- the term frequency of "sing*" and "song*" (which includes "singing", "songs," etc.)
- Potential error: "singed" (not "sang", because Tolkien used antiquated spellings) also meaning burnt.
- Run on all 5 different texts.
- Using tf along with idf, i.e. tf-idf
- Problems with tf-idf: frequency isn't always the best measure of how important a word is in a text, especially fiction
- "The oscillation of who names Luthien – herself or Morgoth – reveals that J.R.R. Tolkien's development of Luthien's agency is not always a steady progression." I wonder if there are other direct passages of text that we can quote that show nearly exact same phrasing of words. That can help us find the "essential differences" between drafts' portrayal of the character, aside from song and dance alone.
- Fuzzy matching using n-grams and regex, using several key scenes (climaxes) that occur repeatedly throughout the 5 texts, to identify what words changed:
- Beren's first encounter with Lúthien
- Lúthien putting guards to sleep with song
- Lúthien's battle against the monster Morgoth
- The naming of Lúthien (herself or Morgoth)
- Method here is similar to those described in Shmidman et al.'s paper identifying similar parallel passages across various versions of the Talmud [2].

2.3.Fuzzy matching with n-grams

Table 1 Editions in More's lifetime

	Editions → / Prefatory material ↓	Louvain: Martens, 1516	Paris: Gourmont, 1517	Basel: Froben, 1518 March	Basel: Froben, 1518 November	Florence: Giunta, 1519	Louvain: Sassen, 1548	Cologne: Birckmann, 1555	Basel: Episcopi 1563
I	Map	1 (1516 map)		4 (1518 map)	4 (1518 map)				10 (1563 n
II	Utopian Alphabet	2		5	5				
III	Utopian poem a) In Utopian b) In transcription c) In Latin translation	3		6	6		5 (only c.)	2 (only c.)	6 (only c
IV	Anemolius' hexastichon	4	1	3	3		4	1	5
V	Giles to Busleyden	5	3	7	7	2	6	7	3
VI	Paludanus to Giles	6	4						
VII	Paludanus' poem	7	5						
VIII	Noviomagus' poem	8	10	11	11	5	10	3	7
IX	Grapheus' poem	9	11	12	12	6	11	4	8
X	Busleyden to More	10	9	10	10	4	9	10	11
XI	More to Giles I	11	6	8	8	3	7	8	4
	Book I	After 11	After 6	After 8	After 8	After 3	After 7	After 8	I: After
	Book II								II: After
XII	Marginal notes	12	7	9 *Helvetii* note omitted	9 *Helvetii* note omitted. Three notes added.		8	9 1518 N included	9 1518 N
XIII	Colophon	13	13	13		7	3	6	2
XIV	Printer's device			14	13	8			
XV	Budé to Lupset		2	2	2		2	5	1
XVI	More to Giles II		8				1 Privilege		12 Rhenar Pirckhei
XVII	Errata		12						X Errata
XVIII	Erasmus to Froben			1	1				the end o
XIX	Other material			(After 2nd title page): 15 Rhenanus to Pirckheimer, introducing the epigrams	(After 2nd title page): 14 Rhenanus to Pirckheimer, introducing the epigrams	1 Erasmus to Foxe After 2: trans. from Lucian by More and Erasmus			volum

Figure 2. Variants in the ngrams across the five texts.

2.5 Sources of error

Bring over the lists of word search errors you wrote about in the OneNote notes of Lay of Leithina

Mention the dffeiretn formats of each of the five texts = e.g. Quenta Noldorinwa is an archivist text, Lay is a poem, etc. More specificity in future needed when designing the search terms and computational methods. As C. Tolkien explained it, " After the hasty 'Sketch of the Mythology' (chapter II in this book), the Quenta Noldorinwa was in fact the only complete version of 'The Silmarillion' that my father ever made. Towards the end of 1937 he interrupted work on a new version, Quenta Silmarillion, which extended to part way through the story of Turin Turambar, and began The Lord of the Rings" (CITE THE SHAPING OF MIDDLE-EARTH).

3. Mixed-method reading

In this section, I bring close reading to bear with the distant reading conclusions we have made above.

4. Postludium

The false end of Tolkien studies (it has been echoed by many researchers that Tolkien studies has been finished, that every detail has been torn apart and repieced together a thousand times. Tolkien himself said that he has 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. What DH can do.

Works cited

- ➔ A song of greater power
- ➔ Women in the works and life of J.R.R. Tolkien by Again Cami
- ➔ Shmidman et al?
- ➔ Lee, J. A Computational Model of Text Reuse in Ancient Literary Texts. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:472-479
- ➔ https://dl.acm.org/doi/10.1145/3195727
- ➔ The Power of music in the tale of beren and luthine by jr.r.r tolkien
- ➔ Your journal article, *towards …*
- ➔ *Your website*
- ➔ *Tauber*
- ➔

Need about 10 more references

[insert Mendeley works cited here] check if its mla

Moore, C. (2021). A Song of Greater Power: Tolkien's Construction of Lúthien Tinúviel. *Mallorn: The Journal of the Tolkien Society*, *1*, 6–16.
Ferré, V. (2021). The Red Book and Tolkien's "The Lord of the Rings": A Fantastic Uncertainty. *Mallorn*, *1*, 26–33.

Genette, G. (1997). *Paratexts : thresholds of interpretation*. Cambridge University Press.
Shmidman, A., Koppel, M., & Porat, E. (2016). *Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus*. https://doi.org/10.46298/jdmdh.1388

Cite Uricchio