Towards Tolkien research with computational literary analysis

Author biography

Christina Dinh Nguyen is from Toronto, Canada. She completed her undergraduate degree in English literature at York University. During her studies, she worked as a research assistant under Professor Anne MacLennan, studying the impacts of Canadian Radio-television and Telecommunications regulations on community, campus, and indigenous radio stations. She is currently a Master of Information student at the University of Toronto. Her research interests are in computational literary analysis, Tolkien studies, and information-seeking behaviors of academics.

Abstract

Since the advent of the term “distant reading” two decades ago, we have seen computational literary analysis give rise to “mixed-method[s] reading” in an attempt to balance out the weaknesses that pure distant reading brings. Computational literary analysis has been spreading across the major fields of literary analysis, including Shakespeare studies and Jane Austen studies, yet there is hardly a peep about it being used for Tolkien studies. I make the case for such research to be done, drawing examples from other fields of computational literary analysis to explain what techniques could be useful here, and why. I also identify potential sources of error and argue for open-source data, in an attempt to encourage methodological transparency and to foster communication between those who prefer close reading and those who prefer distant reading. The appendix includes a literature search log.

Rationale

To sociologists, using computers to study large amounts of digital natural language (e.g. through RSS feeds from social media) is widely accepted. To linguists, using computers to study patterns in speech and written text is also well-established. Programs like Voyant and Gale Digital Scholar Lab can be used to identify, isolate, and describe patterns such as repeated word choice and common grammatical structures within any given corpus. Likewise, English literary studies has recently developed the sub-field of computational literary analysis (CLA), harnessing the power and flexibility of computers to support our scholarly reading of nearly any text. In the areas of poetry and Shakespearean plays, much has been done to study texts through sentiment analysis, word frequency, and topic modelling. Yet surprisingly, there is much less research to be found in the intersection between fantasy and computational literary analysis, and in particular, in Tolkien studies. In this literature review, I consider the two worlds, the world of computational literary analysis and the world of Tolkien literary studies, to highlight what each can gain from the other. In particular, I describe here the *qualities* of Tolkien’s fictional works that are ripe for computational literary analysis, and provide insight on where this area could grow in the near future.

Before we start, I will define the key terms. Precise definitions are scarce in the existing literature, and what does exist tends to nebulously conflict with what others have said. Figure 1 suggests how these terms can fit together under the umbrella of "methods of literary analysis."

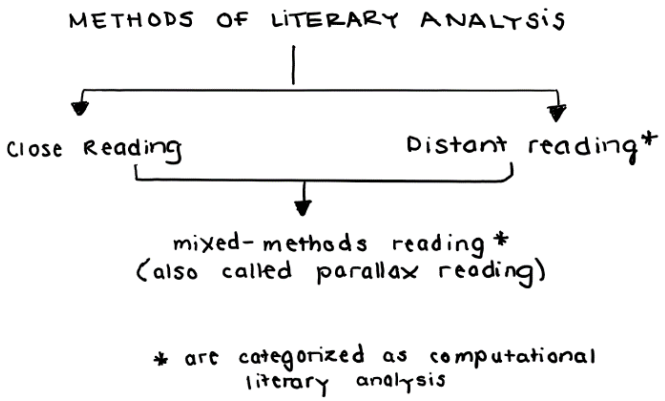

Figure 1. The different methods of literary analysis.

"Computational literary analysis" (CLA) is a relatively new subfield of literary studies, partially popularized by Franco Moretti under the name of "distant reading" in his 2013 book *Distant Reading* (Moretti, *Distant Reading* 54). He originally coined it in a 2000 article (Moretti, *Conjectures on World Literature*). It applies the methodologies of data science and computer science to literary studies. Any type of literary analysis that includes computation is named "computational literary analysis."

"Distant reading" (DR) is the counterpart to the traditional style of close reading. It is computational reading without the benefits that close reading brings: that is, it tends to ignore or forget about context and authorial intent. More will be said about this later.

"Mixed-method reading" (MMR), also called "parallax reading," is the happy mix of close and distant reading. It brings the benefits of close reading, like contextual knowledge, to merge with the benefits of distant reading, like the ability to study large amounts of texts at once for a small, particular detail.

The "corpus" is the text, or section of text, or grouping of texts, under study. In scientific terms, it is the population from which we draw data and create conclusions. It may be a single book, a chapter of a book, a paragraph, a single file with the first chapters from five different books, and so on.

It should be noted that the close and distant reading schools of thought have been critical of each other; close reading scholars complain that distant reading focuses on the statistics too much, not providing an explanation as to the "so what?" question of how it affects how we read whatever we read. Distant reading scholars, on the other hand, often complain that close reading leads to subjective judgements, which should be backed up with clearer (is it implied, statistical) evidence.

The middle line: the mixed-method approach

In recent years, however, some scholars have found the golden middle line. There is room for the strengths of both types of reading to come together: using distant reading's ability to consume large amounts of texts and spot patterns quickly, and our human brain's capacity to recognize the *meaning* of the patterns (both *before* we code, telling the computer what to look

for, and *after* we code, to interpret the computer's data) we gain the mixed-method approach. More about the process of coding is explained in the next section.

Let us look into this golden middle line more, by understanding where we started before reaching the middle. Franco Moretti, the father of distant reading, readily admits the difference between close and distant reading, making the case that

> [In distant reading, the distance] is a *condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, Less is more. If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor
>
> (Moretti, *Distant Reading* 55).

Using large amounts of data can support a hypothesis, Moretti is saying, and we have to accept the cost that that carries, the loss of attention paid to one text, to the context that can only be gained when close reading one text.

Predictably, Nan Z. Da eventually raised an excellent critique of distant reading, and it applies to our reading of Tolkien with CLA (Da, 2019). She argues that

> No matter how fancy the statistical transformations, [CLA] papers make arguments based on the number of times *x* word or gram appears. [CLA]'s processing and visualization of data *are not* interpretations and readings in their own right. To believe that is to mistake basic data work that may or may not lead up to a good interpretation and the interpretive choices that *must be made* in any data work (or have no data work at all) for literary interpretation itself
>
> (Da 606).

There is a real fear that CLA turns literary analysis into too much of a natural science, reducing texts to a series of numbers that tell us much about the *makeup* of a text, but nothing about the *meaning* of a text. Instead, Da says, the research should show that the statistical analysis supports a "so what," rather than just presenting the data *as* the final reading. What use is it to know that there are 459 cases where the word "Gandalf" appears in *The Fellowship of the Ring* (Nguyen)? We can balance Da's critique with Moretti's definition of distant reading to find a happy middle line that brings us the benefits of both their statements. Distant reading is not a sin – unless it is accomplished alone; it *must* be paired with a close reading and a clear 'so what' statement. Parallel to this discussion, readers may wish to read Jay Jin's *Problems of Scale in 'Close' and 'Distant' reading* to understand how we can change how "close" and "distant" reading is defined, in order to better address the interactions between the two fields, and to adjust research methodologies to account for all the weaknesses of where both fields meet (Jin).

Take the example just given about how many times "Gandalf" appears in *FOTR*. With a close reading tacked onto the distant reading, we may say that the 459 cases, when compared to the number of times that "Gandalf" appears in other Middle-Earth stories, tells us a lot about his importance as a character in the plot. Naturally this is an opinion that needs to be fully fleshed-out with contextual readings before being presented as a hypothesis. Consider this also: much of our close reading of Tolkien so far has been based on contextual understandings: for example, the keen reader will recognize the plot parallels between *Children of Húrin* and the medieval Finnish tale *Kalevala.* That is to say that a researcher using CLA to study Tolkien with a new light *must* have already done a close reading of that subject first.

For example, it is not possible for the researcher to study *Children of Húrin* by skipping the close reading and going straight to R to code a program, asking it to find parallels (direct quotations that match, for instance) with other texts. The researcher must know exactly what the two texts are to compare: *CoH* and *Kalevala*. That only comes with close reading. So the mixed-method approach is really the only way forward. Sculley and Pasanek also make some great suggestions on keeping the mixed-method accountable: "[The best practices] for making results from data mining in the humanities domain as meaningful as possible [are to] include methods for keeping the boundary between computational results and subsequent interpretation as clearly delineated as possible" (Sculley and Pasaneck 409). This way, other researchers responding to the CLA can use the same computational data to draw varying conclusions, strengthening the academic conversation by allowing discourse to flow clearly. Later on, I also explain why publicly-accessible code and computational data is important in the computational literary analysis field.

"Great Literature," i.e. the stuff of traditional literary canon, has been experiencing a shakeup in analysis methodology in the past three decades. Some notable examples are Shakespeare studies and studies into Sylvia Plath's poetry. Shakespeare's plays and his poetry alike have benefitted from computational literary analysis. Researchers like Mireille Ravassat and Jonathan Culpeper have employed transdisciplinary methods (in Ravassat and Culpeper's case, they used methods from linguistics, statistics, and literary studies) to uncover a new vision of how we read Shakespeare (Ravassat and Culpeper). Shakespeare was perhaps as much of a skilled linguist as Tolkien. He enjoyed the rhythm and beat of his words (or perhaps that is how it appears to us, the hearers of his words) – linguists have long used Shakespearean texts as corpuses in their studies, and so have literary analysts (Ravassat and Culpeper). So, after all, why should this not be done for Tolkien, who, aside from reviving archaic English words in his Middle-Earth stories, also invented whole languages like Qenya? Words like "oliphaunt," "pipe-weed," and "mithril" deserve to be tracked: where and when do they occur? In what density do they make up the texts? Indeed, how much of *The Lord of the Rings* is made of invented words; how much is made of Latinate words, and how much is made of Germanic words? Can we compare this ratio to his later works, like *Morgoth's Ring*? What can this comparison tell us about the evolution of Tolkien's word-choices over time, and how does it relate to the plot?

Sylvia Plath's poetry has also been analyzed with CLA. Wadsworth et al. found that there were potential correlations between events in Plath's life and changes in her textual sentiment

(Wadsworth et al.). For example, they saw that "[t]he period in 1959, which is here cited as a quantitative shift in style, includes the conception of Plath's first child (June 1959)" (Wadsworth et al. 668). Of course, there is only so much we can use as statistical "evidence" of her decline during the birth of her first child; it is easy to say that the birth caused the sentiment change, but it is impossible to prove (and neither should we assume that the narrator is the same voice as the author). But still, having the statistical data is important; we have the keys *of* understanding Plath and her mental framework, but not the keys *to* Plath. Likewise, sentiment analysis and topic modelling tests can be performed on Tolkien's fictional works. What is he writing most about in *The Hobbit*: the idyllic Shire or the hellscape of Mordor? How much does the sentiment value change between the two extremes of -1 (extremely negative) and +1 (extremely positive); i.e. what is the slope of the sentiment graph? What can this tell us about the tension of the plot, and how can we visualize the shape of the plot?

There are also many tools in the toolbox for mixed-method reading, especially in the distant reading step. Researchers can use pre-made applications like Voyant or the Gale Digital Scholar Lab, which provides a highly interactive, easy-to-understand user interface with many options for distant reading tests (Miller). With these pre-made applications, data is clearly organized, and running distant reading tests are straightforward. However, with make-it-yourself scripts like using R or Python, humanities scholars are afforded a different, more intimate view of the project, and ultimately, there is a greater flexibility in the types of tests that they can run, limited only by the scope of their coding ability. There are also make-it-yourself programs that are based on other peoples' scripts, which gives a researcher a tree trunk to put their leaves on, so to speak. One great example of this in action is in Lundy's study which models the topics of 1,000+ bestseller books; Lundy uses the Machine Learning for Language Toolkit (MALLET) program, which is Java-based, and adapts it to their own needs when topic modelling. The source code, like most of the make-it-yourself scripts based on other peoples' programs, is on GitHub (University of Massachusetts-Amherst).

More qualities of Tolkien suited for CLA

Furthermore, even if the average human brain could remember all the names of all the wars, characters, locations, v.v. of the Eä universe, keeping track of them in such a way as to recognize patterns is virtually impossible. I heartily accept that the only person who has ever kept them straight was probably the Professor himself – and even he admits often in his letters that these wars, characters, and locations are frequently edited, changed, updated, or removed altogether (Tolkien, *The Letters of J.R.R. Tolkien* 127). Furthermore, many of them were based on other ur-texts; Elizabeth Solopova and Stuart Lee, for example, provide us some keys of Middle-Earth from medieval texts that Tolkien read (Solopova and Lee). Note that they argue that the keys are *of* Middle-Earth, not the keys *to* Middle Earth; this is an attitude that we should take up in the field of CLA too – we cannot presume to *prove* that *y* is true because there is statistical evidence of *x;* instead, we should say that *y* is likely true, or may be true, because we have observed *x,* which is known to correlate with *y* (and we should not forget that correlation does not always mean causation).

Returning to the point about the complexity of the Eä universe: there is simply an overwhelming amount of information to process. Brian Kokensparger acknowledges this phenomenon of "linguistic complexity" and information overload with his experience of reading Shakespeare: "[Shakespeare's] language got in the way of the experience. Yet I felt that the effort was worth the gain: a whole new world of literature was opened to me" (Chen 183). If we can put the effort into understanding the language, the *whole* of the linguistic complexities across Tolkien's works, so too will a whole new world of literature be opened to us.

This overwhelming amount of information to process shows where computational literary analysis can help: the human predicts a pattern, then runs a test to confirm that pattern. It is important to note here that the tests are run *after a hypothesis* has been formed – the computer is not doing everything!

There are two types of intertextual readings that I can immediately see using CLA. The first are intertextual references to "outside-of-Tolkien's-writing" sources; the second are intertextual references that are "in-universe" for Tolkien. This latter includes references to previous wars, people, major events, or languages.

There are many intertextual references present in Tolkien's Eä, mostly to medieval sources. These sources came from multiple languages, including (but not being limited to) Old English, Middle English, Finnish and medieval Latin. I have already mentioned the relationship between *Children of Húrin* and *Kalevala,* which Tom Shippey analyzed (Chance). Using text mining techniques already established by others, we too can look for snippets of similar text across a large corpus, particularly across several texts supposedly telling the same story in different variants (the ur-text and its children). For example, Shmidman, Koppel, and Porat used this technique for studying 'parallel passages' (as they called those deviations of the same story) across a large Hebrew and Aramaic corpus. They looked at words' spatial relationships with other words and compared that data between different texts. To be specific, they were "finding matched pairs of strings of four or five words that differ by at most one word and then identifying clusters of such matched pairs. Using this method, over 4600 parallel pairs of passages were identified in the Babylonian Talmud, a Hebrew-Aramaic corpus of over 1.8 million words, in just over 11 seconds" (Shmidman et al. 1). This same method (called "fuzzy matching") could be used with varying number-of-word boundaries to see how similar *Children of Húrin* and *Kalevala* are: are there specific phrases that are nearly the same in both? Are the adjectives used in both cognates of each other, with the understanding that *CoH* was written in English and *Kalevala* was in Finnish? How does the specific translation of *Kalevala* we use affect our data? When we have identified parallel passages, can we identify how the story has changed and how that has affected the location of the climax, or character development?

Shmidman et al. also verified their computational findings by performing empirical comparisons on sample data to check that "the coverage obtained by [the computational method] is essentially the same as that obtained using slow exhaustive methods" (Shmidman et al. 1). In other words, the computer can replace close reading done by humans when it comes to finding parallel quotations – but we also have to do random spot-checks to ensure that the parallels the program picks up are legitimate. We must ensure that when using CLA in Tolkien studies, we

perform such empirical checks too, to verify the usefulness and efficiency of the computational method we have created. We must identify potential sources of error, and if possible, count for those when creating our methodology. Should that prove impossible, then we should clearly explain these potential sources of error in our research findings. An inefficient computational method does not mean that we should discount CLA entirely, either; instead, we should adjust the parameters of our tests to ensure a more accurate identification of parallel passages, such as studying strings of five words at a time instead of four.

Book historians too will find CLA tools useful for understanding textual variants and mapping the interior of a book. Matthew Kirschenbaum and Sarah Werner note in *Digital Scholarship and Digital Studies*, "the instability of texts is a regular feature throughout textual transmission histories," so using tools like high-resolution scans of Tolkien's manuscripts, alongside transcriptions with metadata and hypertext, can only enhance our understanding of how the stories evolved as Tolkien wrote and re-wrote them over time (Kirschenbaum and Werner). After all, Tolkien was often intentionally "plac[ing] his [works] in the same manuscript tradition as other mythologies," with variations on the same characters and events (Moore); it practically demands that we study these textual variants closer, as Clare Moore has done with the Lúthien Tinúviel story (though her methods do not use CLA).

Then, of course, the linguistic aspects of Tolkien's works bear mentioning. Computational linguistics is alive and kicking, often using Tolkien's work as a corpus. Just take a look at Barnes' paper, *Virtual languages in science fiction and fantasy literature*, for a good starting spot. Since linguistics is outside the scope of this literature review, however, I will refrain from further comments.

The wearing of two caps

Let us backtrack a little. I previously mentioned that we, the human, have to understand what we are looking for *before* and *after* we write the code. The CLA researcher has to be able to 'wear two caps,' as it is: the literary analysis cap and the computational researcher cap. They must understand what their argument about the text is, what method would traditionally (i.e. in close reading) used to support that argument, and what that method would look like in computational literary analysis. In other words, they cannot always depend on getting the numbers first, then creating an argument from the numbers. As Kenneth Ward Church quotes in *Emerging Trends*: "Numbers offer the sheen of objectivity; algorithms seem to 'transcend morality', as O'Neil put it, when in fact they only obfuscate the human assumptions that go into creating them" (Church 474).

A hypothesis formed from close reading should be the first step in every research project in CLA. Let us take a small example from *Leaf by Niggle*, and walk through it step-by-step. The source code can be found here via GitHub, for readers who would like to follow along: shorturl.at/kBCJ3.

1. Close reading. *Leaf by Niggle* is well-known for being an allegory of a Christian journey through life and into heaven. A person reading *Leaf by Niggle* for the first time might not have picked this up. They must have performed a close reading, reading it multiple times

and potentially having to research the Christian faith and notions of the afterlife, before fully appreciating the allegorical qualities of *Leaf by Niggle.* They would also read some scholarly literature on *Leaf by Niggle* such as Marie Nelson's *An allegory in transformation* (Nelson). Now they have an idea of what their argument might be. In this case, my argument is that Tolkien decreases the importance of the painting as the story progresses and increases the importance of the tree as the story progresses. It is only after the researcher has done this prefatory work that they can think about the computational literary analysis.

2. Computation. Now I, the researcher, can translate that argument into something a computer can test for evidence of. Let us ask R to track the occurrences of "painting" and "tree," and mark down exactly when in the book it happens. Creating a dispersion plot, or a "token distribution plot," will help us *visually* determine when in the story the words "picture" and "tree" tend to occur. The further along in a text a word is, the more time has passed. Therefore, we will call the x-axis of these plots "novelistic time" and the y-axis will be the YES/NO occurrences of such words. There are no numerical values for the y-axis since a black line indicates the YES, and a blank/white line indicates a NO.

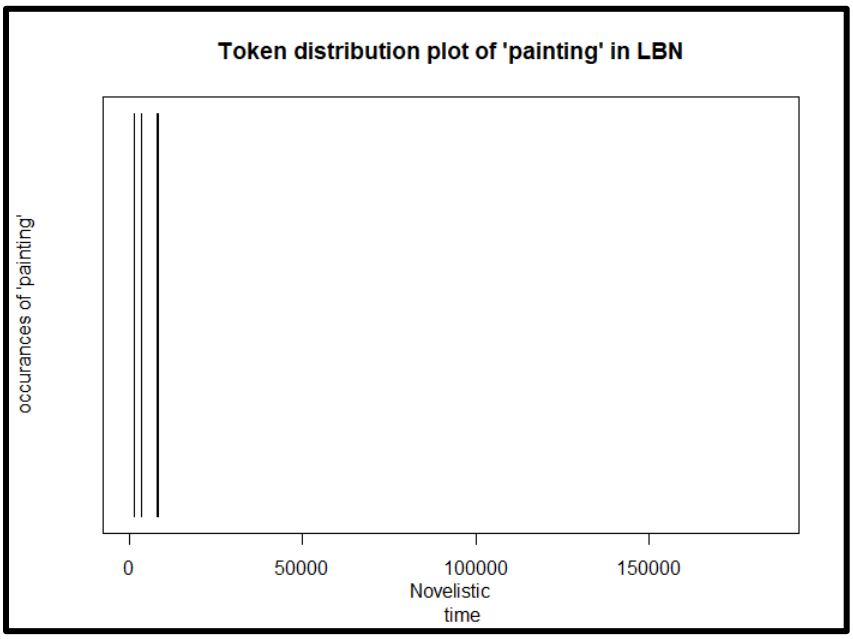

Figure 2. Token distribution plot of 'painting' in 'Leaf by Niggle.'

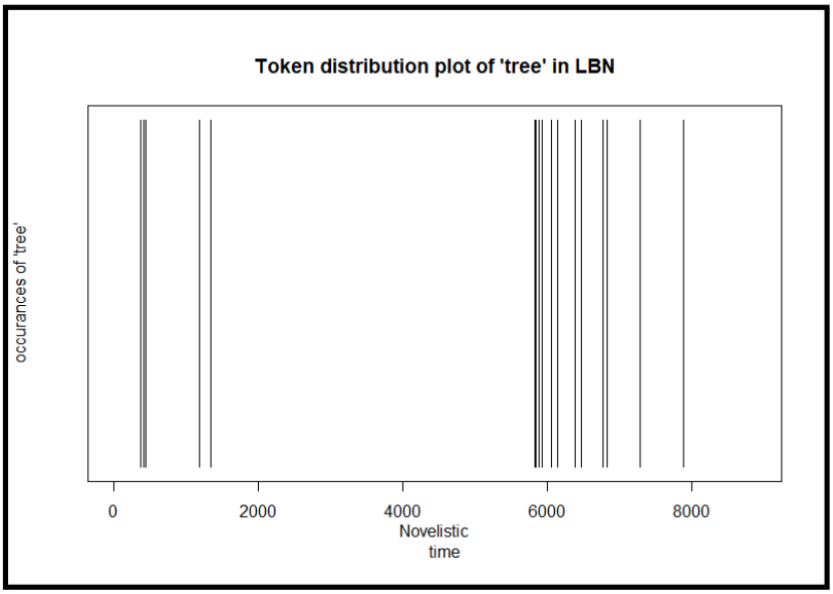

Figure 3. Token distribution plot of 'tree' in 'Leaf by Niggle.'

It looks like there is not a very strong relationship between *painting* and *tree* (there is not enough data to make a conclusion). I go back to re-read the text to see if there is a different word that Tolkien uses the describe the painting more often. There is, it is *picture.* So I re-run the same test, but I ask R to compare *picture* and *tree* instead of *painting* and *tree.* This step shows why it is important that I do the close reading when performing CLA. Here is what shows up:

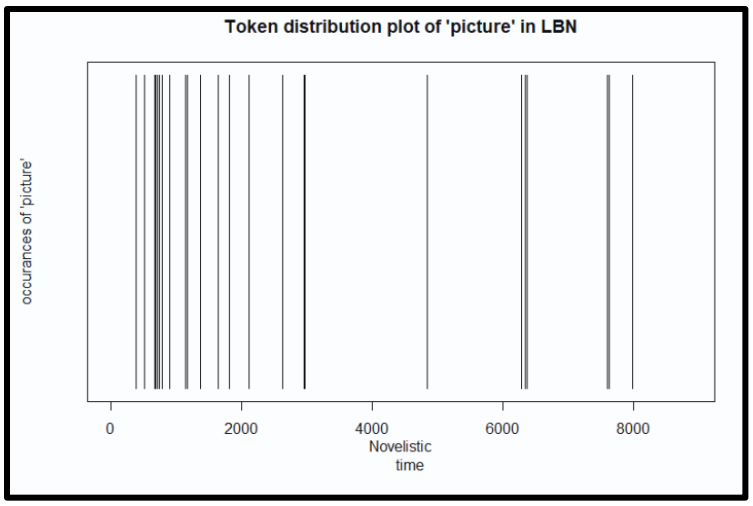

Figure 4. Token distribution plot of ‘picture’ in ‘Leaf by Niggle.’

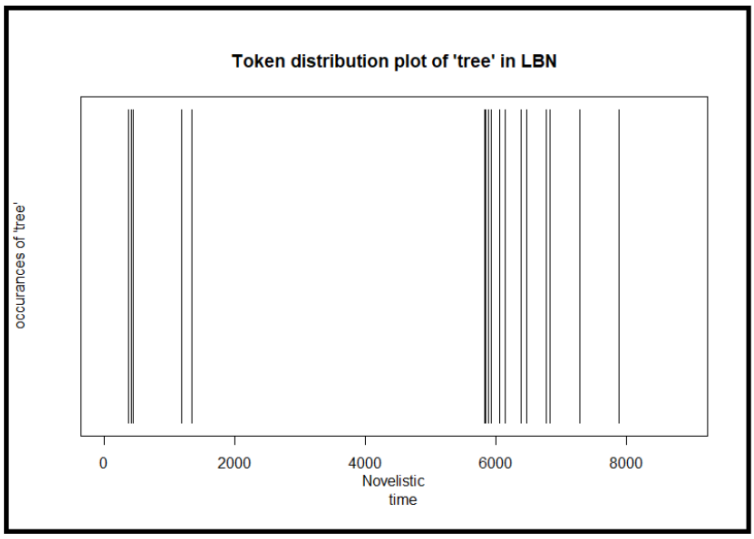

Figure 5. Token distribution plot of 'tree' in 'Leaf by Niggle,' printed again.

3. Literary analysis. We can clearly see an inverse relationship: as the number of "picture*"* goes down, the number of "tree" goes up. An argument could be made that when picture becomes less important to the plot, tree's importance rises, and so it is mentioned more and more as Niggle encounters the realized vision.

4. Identifying potential sources of error. I made a major assumption when plotting the occurrences of these tokenized words: that the further along in a text we are, the further along in time we are. This is patently untrue in many cases; take *The Two Towers*, for example. After the fellowship of the ring split into smaller groups to continue their quests at the end of *The Fellowship of the Ring*, we have back-to-back chapters that describe events happening at the same time (or even having later chapters describing events from further back than that), but to different people. Therefore, the events of the chapter called "The Uruk-Hai" should not be confused as happening after the events of "The Riders of Rohan," though the "The Uruk-Hai" chapter appears directly after the "The Riders of Rohan" chapter.

The ways forward

As I mentioned in the introduction, other Tolkien texts could benefit from parallax reading. Long stories, like those in *The Lord of the Rings*, are an obvious first choice, because of the sheer quantity of text to be processed and analysed. For example, it certainly would not do to sit and highlight all the Germanic or Latinate words in *LOTR* by hand, nor to count them manually. However, a program in R, using tools like <quanteda> or <wordcount>, could do very well. The less obvious corpus choices for parallax reading, however, like *Leaf by Niggle*, are short stories. Trends are easier to spot when close reading, erroneous distant reading conclusions can be quickly verified by a second or third close reading.

Another way that Tolkien studies could benefit from computational literary analysis is less literary and more linguistic. Johns and Randall showed in their research how distant reading can help identify patterns in large corpuses of text that allow us to pinpoint *where* and *when* the authors were writing (Johns et al.). This application is related to machine learning, where we can train a program on many texts *with* their corresponding metadata, then ask the machine to read a sample and estimate where and when it was written.

There are some glaring potential pitfalls and weaknesses too, which I hope we can avoid by listing them here. I have already mentioned the obsession with distant reading (statistics) while avoiding the mixed-method reading. We also need to eventually develop a centralized corpus, preferably open source, of all of Tolkien's works, including his fiction, his letters, transcripts of his lectures, and vice versa. Having this centralized corpus will remove the potential for biased subsets and will further encourage all Tolkien studies researchers who perform CLA to use the same data. As Dr. Stuart Lee mentioned in his lecture, *The Digital Researcher and the humanities*, one of the greatest challenges for future digital humanities researchers is recognizing that "data for analysis needs to be readily available and open," in a manner copying what we have in the physical sciences: this will increase the reproducibility of published results, which is of paramount importance to any statistical or scientific research (Lee). That is, your truth is the truth that others can find, too. In this way, you make your experiment transparent and allow others to grasp what was done, whether or not they will go on to replicate your results.

One great example of a fully transparent project is Sparrow (Laurie Frances) Alden's *Words that you were saying* (Alden). In it, Alden logs "pleasing patterns of wonderful words in *The Hobbit*," including diction changes, archaic terminology, gastronomy terms, and so on. Her results are wonderful sets of visualizations (e.g. graphs of narrative time vs. use of archaic terms), a result of various Python scripts, which are open-access via a GitHub repository, posted by one of her children (Alden). Both the raw and processed results are available directly on the website, typically as PDF links, and so other scholars are invited to try processing the data on their own. What *Words that you were saying* lacks, however, is a standardized format for presenting the raw and process data; instead, Alden has blog posts describing her results, and the PDF format is not the most conducive to statistical analysis (for example, text mining in R will typically use .csv or other tabular data formats). This blog style is eminently useful for passing visitors curious in the outcomes, but it does make replicating Alden's work a tad bit more difficult. This is evident when looking at several scholar's independent project websites on Tolkien texts (that datasets are decentralized). Thus again I repeat that having a standardized, centralized dataset for computational literary analysis of Tolkien is important – the raw data (and some semi-processed) should not be spread out over the websites of several scholars. Other considerations for this standardized dataset should include accessibility to multiple programming languages (Python and R being the largest contenders that come to mind) and multiple programming packages (for example, R text mining often uses <quanteda>, <tidytext>, <ggplot2>, and <tm>, some of which do not work tidily with the others).

Another project dealing with computational literary analysis and Tolkien's fiction is James Tauber's *Digital Tolkien*. Being an on-going project rather than a master's thesis (as Alden's was), Tauber's focus is considerably broader; as of the time of writing, he is working on many subprojects dealing with many of Tolkien's works, not being limited to *The Hobbit* or indeed even Middle-Earth. Tauber makes great strides in laying out the groundwork for future scholars; he has applied (and is continuing to apply) many of the major techniques in text mining to those texts, including visualizing textual variants and cataloguing verse with rhyme and meter. He has *not* provided what Nan Z. Da calls the "so what;" instead his work is mostly data work without the close reading, or mixed-method reading, element. This distant reading can support mixed-methods reading of future scholars looking for "hard" evidence of some spotted trend, or to reject previous arguments. Tauber's use of an XML TEI (Text Encoding Initiative) tagging format is also in-line with current digital humanities trends, helping standardize parts-of-speech tagging. However, while Tauber's project is steadily trotting along, again this is a good example of raw and semi-processed data being spread across multiple scholars' sites. This is not a fault of the individual scholar, I should emphasise; it would benefit *all* digital Tolkien scholars to share and collaborate on data-gathering and processing expeditions – this avoids unnecessarily doing the same task twice, and it will increase transparency as well as efficiency.

From a library and information sciences perspective, the other major attraction of open-access digital humanities is the reduction of information poverty and the increase of information democracy. Information poverty is defined as a scenario in which individuals and groups lack the necessary skills, abilities, or material resources to acquire efficient access to information, interpret it, and apply it correctly in a particular environment. Much of the arguments about open-access data are tangential to the nature of this paper, so I will only suggest that the reader looks at other works about open-access data, such as Peter Suber's excellent article, *Promoting Open Access in the Humanities* (Suber).

Because of the nature of Tolkien studies where many drafts of a single story were often written, that standardized corpus should account for this too. Here we can learn from Gerlach and Font-Clos's proposed Standardized Project Gutenberg Corpus, which attempts to collect all of Project Gutenberg's texts (which have commonly been used for linguistic and literary analysis, yet in a disjointed, decentralized manner) into a single corpus that is standard across the field of computational linguistic analysis (Gerlach and Font-Clos). Some meta-data should also be added to the standardized Tolkien corpus, such as time of publication, publisher, languages involved, v.v. which will help researchers sort the texts quickly by characteristic. The standardized Tolkien corpus could, in turn, be expanded to include secondary sources like the works that Tolkien used for inspiration, like the *Kalevala* or *Beowulf.* This expanded standardized corpus could make use of Shmidman et al.'s (Shmidman et al.) techniques to discover parallel text fragments, and then be incorporated into larger CLA projects like the Common Language Resources and Technology Infrastructure (CLARIN). CLARIN provides the texts (to use as corpuses) and tools to analyse those texts; at the time of writing, CLARIN has data for multiple topic areas and spans many languages, with a Euro-centric approach. The proposed Tolkien dataset certainly belongs here with the other digitized versions of Great Literature.

Aside from disjointed data sources, the field of Tolkien studies and CLA will certainly be using "different filtering techniques to mine, parse, select, tokenize, and clean the data," as in the case of computational linguistics (Gerlach and Font-Clos 2). This diversity of methodology is undoubtedly a strength, since differing techniques will provide different perspectives that, when combined, will provide a larger coherent picture of our subject of study. But Gerlach and Font-Clos also point out that there must be clear descriptions of the methodological steps in "sufficient detail," so that such combined pictures *can* be possible (Gerlach and Font-Clos 2). If one study uses a systematic review, and another uses a scoping review, how can we combine the conclusions without first seeing the different methodologies?

Two methodologies in particular seem to be a good starting point for Tolkien studies with CLA: identifying cognitive style features and lexical features. In *Cognitive Learning*, Liu et al defines cognitive style features analysis, which studies "the proportion of sentences containing 1st person, time, certainty, and certainty indicators" (Liu et al. 3). Lexical features, on the other hand, "captures one's characteristics in word usage" (Liu et al. 3). By providing a baseline of cognitive style features and lexical features of Tolkien, especially of his fictional works, we can get an idea of what it means to "sound like Tolkien" or to be "Tolkien-esque"; that is, to use certain types of words in the same proportions and contexts that Tolkien did. This baseline could also help us consider how his cognitive style changed over his career: for example, did the ratio of Germanic:Latinate words change over time? How can we correlate this to the books he might have been reading during that time of change?

"Methodologies" also implies using different programming languages. Currently, Python and R are immensely popular for data mining and computational literary analysis, but there are many more tools and languages in the field (Leetaru). There must be a way to account for different languages and how they affect the processing and analysis of the data.

Conclusion

I am optimistic about the possibilities that computational literary analysis brings to Tolkien studies, both in reaffirming past research and pushing the boundaries of new research. It comes altogether as a surprise that little work has been done in bringing distant reading or mixed-method reading to Tolkien. As I have highlighted with the examples of previous research, scholars from various fields need to come together to tackle the interdisciplinary nature of these projects.

There is also ample space in both the fields of Tolkien studies and digital humanities for a discussion of how to use all the methodologies described above to suit the needs of particular projects; not all the tools of natural language processing in other genres carry over to fantasy. Furthermore, there is a need for more applied research of computational literary analysis of Tolkien works, along with mixed-method approaches that allow for the voices of both close and distant reading to maintain a dialogue with each other. It is in this way that I hope we can continue to breathe life into the stories we love, and see that "still round the corner there may wait / [a] new road or a secret gate" to be discovered (Tolkien, *The Fellowship of the Ring*).

Appendix

Here is the literature search tracking log used for this literature review, when finding resources on the intersection between Tolkien studies and computational literary analysis.

Search #	Database or search engine	Years searched	Boolean	Number of hits	Notes/pulled which sources?
1	OneSearch, University of Toronto Library	1950--	(tolkien OR fantasy) AND ("computational literary analysis" OR "computational analysis" OR "distant reading ")	712	Results are not specific enough to capture Tolkien studies or fantasy studies. However, there are many sources for distant reading; I am pulling some of those; critiques of distant reading, applied cases of computational literary analysis, and some theory of distant reading
2	OneSearch, University of Toronto Library	1950--	(tolkien OR fantasy OR English literature) AND ("computational literary analysis" OR "computational analysis" OR "distant reading " OR "text mining") Articles only	8, 818	This search captured more critiques of distant reading and many theory-based perspectives on how distant reading can be applied to literature. I am pulling some of these sources to balance out the literature review with pros/cons of distant reading.
3	OneSearch, University of Toronto Library	1950--	(shakespeare OR bible OR poetry) AND ("computational literary analysis" OR "computational analysis" OR "distant reading " OR "text mining")	497	This search capture some great applications of distant reading *and* mixed-method reading, which gives me a fleshed-out picture of how those two methods differ in approach and how mixed-method reading might be better (more accurate) for humanties research than distant reading. I am pulling some of these sources to use as examples to draw on for the lit. review, e.g. "we can use _ paper's model to do similar tests on Tolkien literature."
4	OneSearch, University of Toronto Library	1950--	(aramaic OR hebrew OR torah) AND ("fuzzy reading" OR fuzzy OR computational literary analysis OR distant reading) Articles only, English language only	15, 542	This search gave too many hits to be very productive, but it did confirm that there was some research out doing distant reading on Jewish theological texts. I need to be more specific, or get fewer results, e.g. by trying a different, more specialized search engine
5	arXiv	N/A	(aramaic OR hebrew OR torah) AND ("fuzzy reading" OR fuzzy OR computational literary analysis OR distant reading)	1	This search gave me 1 hit, which was a great paper that showed how applied computational literary analysis was done to a corpus of Hebrew poetry. Title: "ViS-Á- ViS : Detecting Similar Patterns in Annotated Literary Text," arXiv:2009.02063.
6	arXiv	N/A	(parallel passages) *in abstract* AND (torah OR bible OR old testament OR hebrew OR aramaic) *in title*	2	This search gave me 2 hits, which were both great papers that showed how applied CLA was done to a corpus of Hebrew texts. I am pulling both of these to read for the literature review, similar reasoning as search #3 above.
7	arXiv	N/A	(distant reading OR parallel passages OR text mining) *in abstract* AND (literature)	2	This purposefully broadened search only gave me 2 hits which were not very related to what I need, so I will move away from arXiv to see what other data- bases might work.
8	ProQuest ERIC	1950--	(tolkien OR fantasy OR english literature) AND ("computational literary analysis" OR "computational analysis" OR "distant reading ")	3	These hits were not useful since they dealt with genres too far away from either 'Great Literature' or Tolkien and fantasy studies.
9	ProQuest ERIC	N/A	(tolkien OR fantasy OR English literature) AND ("computational literary analysis" OR "computational analysis" OR "distant reading " OR "text mining") Articles only	297	There are many useful hits of applied CLA, especially to the KJV Bible and other corpuses of 'Great Literature.' I am pulling these to read and potentially use as examples to draw on in the literature review.
10	ProQuest Linguistics and Language Behavior Abstracts	N/A	(tolkien OR fantasy OR english literature) AND ("computational literary analysis" OR "computational analysis" OR "distant reading " OR "parallel passages") NOT ("ESL" or "secondary language")	5, 952	While there are an overwhelming number of hits, the first few hits are theory-based papers that will be useful in understanding what the different tools of CLA are. There is a lot of discussion about bringing in practices from other discplines and making the research truly interdisciplinary.

References

Alden, Laurie Frances (Sparrow). *Words that you were saying*. n.d. 17 January 2022. <https://wordsthatyouweresaying.blog/about/>.

Barnes, Lawrie, and Chantelle van Heerden. "Virtual Languages in Science Fiction and Fantasy Literature." *Language Matters (Pretoria, South Africa)*, vol. 37, no. 1, University of South Africa, 2006, pp. 102–17, doi:10.1080/10228190608566254.

Birns, Nicholas. "The Children of Húrin, Narn i Chîn Húrin: The Tale of the Children of Húrin (Review)." *Tolkien Studies*, vol. 5, no. 1, West Virginia University Press, 2008, pp. 189–200, doi:10.1353/tks.0.0022.

Chance, Jane. *Tolkien and the Invention of Myth: A Reader*. University Press of Kentucky, 2004.

Chen, Shu-Heng. *Big Data in Computational Social Science and Humanities*. 1st ed. 20, Springer International Publishing, 2018, doi:10.1007/978-3-319-95465-3.

Church, Kenneth Ward. "Emerging Trends: I Did It, I Did It, I Did It, But." *Natural Language Engineering*, vol. 23, no. 3, Cambridge University Press, 2017, pp. 473–80, doi:10.1017/S1351324917000067.

Culpeper, Jonathan, et al. "Measuring Emotional Temperatures in Shakespeare's Drama." *English Text Construction*, vol. 11, no. 1, 2018, pp. 10–37, doi:10.1075/etc.00002.cul.

Da, Nan Z. "The Computational Case against Computational Literary Studies." *Critical Inquiry*, vol. 45, no. 3, The University of Chicago Press, 2019, pp. 601–39, doi:10.1086/702594.

Gerlach, Martin, and Frances Font-Clos. "A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics." *Entropy (Basel, Switzerland)*, vol. 22, no. 1, MDPI, 2020, p. 126, doi:10.3390/e22010126.

Jin, Jay. "Problems of Scale in 'Close' and 'Distant' Reading." *Philological Quarterly*, vol. 96, no. 1, UNIV IOWA, 2017, pp. 105–29.

Johns, Brendan T., and Randall K. Jamieson. "The Influence of Place and Time on Lexical Behavior: A Distributional Analysis." *Behavior Research Methods*, vol. 51, no. 6, Springer US, 2019, pp. 2438–53, doi:10.3758/s13428-019-01289-z.

Kirschenbaum, Matthew and Sarah Werner. "Digital Scholarship and Digital Studies: The State of the Discipline." *Book History* 17 (2014): 406-458. 17 January 2022.

Lee, Stuart. "The Digital Researcher and the Humanities." *Dr Stuart Lee - The Digital Researcher and the Humanities - Part 2*. RDTFDiscovery, 2011. Film. 2021. <https://www.youtube.com/watch?v=0XBnILeKTNo>.

Leetaru, Kalev. *Data Mining Methods for the Content Analyst an Introduction to the Computational Analysis of Content*. Routledge, 2012.

Liu, Xiaotong, et al. "Cognitive Learning: How to Become William Shakespeare." *Conference on Human Factors in Computing Systems*, ACM, 2019, pp. 1–6, doi:10.1145/3290607.3312844.

Lundy, Morgan. *Text Mining Contemporary Popular Fiction: Natural Language Processing-Derived Themes Across over 1,000 New York Times Bestsellers and Genre Fiction Novels*. ProQuest Dissertations Publishing, 2020.

Miller, A. "Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools." *Journal of Web Librarianship*, vol. 12, no. 3, Routledge, 2018, pp. 169–97, doi:10.1080/19322909.2018.1479673.

Moore, Clare. "A Song of Greater Power: Tolkien's Construction of Lúthien Tinúviel ." *Mallorn: The Journal of the Tolkien Society* 62 (2021): 6-16. 17 January 2022.

Moretti, Franco. "Conjectures on World Literature." *New Left Review*, vol. 1, no. 1, New Left Review Ltd, 2000, p. 54.

Moretti, Franco. *Distant Reading*. Verso, 2013.

Nelson, Marie. "J.R.R. Tolkien's 'Leaf by Niggle': An Allegory in Transformation." *Mythlore*, vol. 28, no. 3/4 (109/110), Mythopoeic Society, Oct. 2010, pp. 5–19, http://www.jstor.org.myaccess.library.utoronto.ca/stable/26814906.

Nguyen, Christina Dinh. "Visualizing Tolkien with R, Project 6, Step 1 (FOTR only)." 1 December 2021. *GitHub.* R code. 4 December 2021. <https://github.com/TorontoYYZ/VisualizingTolkienwithR/blob/c998468c23f44f725582f104a88b87e0521aa3a1/Project%206,%20Step%201%20(FOTR%20only)>.

Ravassat, Mireille, and Jonathan Culpeper. *Stylistics and Shakespeare's Language: Transdisciplinary Approaches*. New York, 2011.

Sculley, D., and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing*, vol. 23, no. 4, Oxford University Press, 2008, pp. 409–24, doi:10.1093/llc/fqn019.

Shmidman, Avi, et al. *Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus*. 2016, doi:10.46298/jdmdh.1388.

Solopova, Elizabeth, and Stuart Lee. *The Keys of Middle-Earth: Discovering Medieval Literature through the Fiction of J.R.R. Tolkien*. Palgrave Macmillan, 2015.

Suber, Peter. "Promoting Open Access in the Humanities." *Syllecta Classica*, vol. 16, no. 1, University of Iowa, Department of Classics, 2005, pp. 231–46, doi:10.1353/syl.2005.0001.

Tauber, James. *Digital Tolkien*. n.d. 17 January 2022. <https://digitaltolkien.com/about/>.

Tolkien, J. R. R. (John Ronald Reuel), and J. R. R. (John Ronald Reuel) Tolkien. *The Fellowship of the Ring : Being the First Part of The Lord of the Rings* . Grafton, 1991.

Tolkien, J. R. R. (John Ronald Reuel), et al. *The Letters of J.R.R. Tolkien*. George Allen & Unwin Ltd., 1981.

University of Massachusetts-Amherst. *MALLET*. 11 August 2021. 4 December 2021. <https://github.com/mimno/Mallet>.

Wadsworth, Fabian, et al. "Evolution of Vocabulary in the Poetry of Sylvia Plath." *Digital Scholarship in the Humanities*, vol. 32, no. 3, 2017, p. 660, doi:10.1093/llc/fqw026.