Clichés and the Reception of 19th-Century English Novelists

Holly Cui and **Jennifer Ouyang** and **Heidi Smith** and **Amy Weng**
Duke University, Durham, North Carolina
yifan.cui@duke.edu
jennifer.ouyang@duke.edu
heidi.a.smith@duke.edu
amy.weng@duke.edu

Abstract

If we wanted to be cutesy, we could start this paper with a cliché—but that would be like beating a dead horse, now wouldn't it? Following the work of van Cranenburgh (2018) on the presence of clichés in modern Dutch literature versus their literary quality, we wanted to explore whether this phenomenon was replicable in 19th-century English novels. Expanding on the original question, we explored the markers of scholarly acclaim, quality, and popularity based on Goodreads and Modern Language Association data as compared to the percentage of words in an authors' texts from Project Gutenberg that contain clichéd expressions. Our results show that, similarly to the original study, scholarly acclaim has a negative correlation with cliché ratios, whereas the other two predictor variables are positively correlated. We also summarize our findings vis-à-vis the clichés themselves, which are present in our corpus. The code used for this paper is available at https://github.com/amycweng/cliches-and-19th-Century-English-Novelists.git.

1 Overview

By cliché, we are referring to a multi-word expression (MWE) that is frequently used and thus easily recognizable as a unit. MWEs function as a single word, as the one-to-one translation of such expressions does not make sense and thus must be parsed differently than regular words in some contexts (Constant, 2017).

Past research into clichés in various media has used the method of regular expressions to detect either pattern-based variable phrase detection (Sweed and Shahaf, 2021), TF-IDF (Smith et al., 2012), n-gram and rhyming pairs (Smith et al., 2009), or simple fixed pattern matching (Cranenburgh, 2018). Cook and Hirst (2013) performed early research on this topic, exploring methods previously used for contexts such as the "Eumaeus" episode of Joyce's *Ulysses*, in which the protagonist Leopold Bloom drunkenly makes his way home. The author compared n-grams in the episodes of the epic to other works of the same period to determine that it has the most hackneyed language of the novel. This is interesting to our research because *Ulysses* is perhaps one of the most prestigious literary novels of the canon of novels written in English; yet, it exhibits a large number of clichés, purposefully so, which is contrary to computational trends and the general literary opinion.

2 Problem Statement

We will be performing a replication study of the aforementioned paper by Cranenburgh on the prevalence of clichés compared to the perceived literariness and quality of Dutch novels but on 19th-century novels written in English. We will focus on a subset of well-known authors from that period and look at the average quantity of clichés in their works compared to their scholarly and popular reception to test whether this phenomenon is replicable in English literature, specifically that which is readily available in the public domain.

3 Nineteenth Century English Novelists

Unfortunately, we found no English equivalent to the large reader survey that rated 401 contemporary Dutch language novels by their literariness and quality used by van Cranenburgh (2018), and we are restricted to working on older texts available in the public domain. Beginning with a list of 356 writers listed on a Wikipedia category page[1] of 19th-century novelists born in England, we found that 97 of them have works relevant to our study in Project Gutenberg's catalog[2]. A relevant work is an original English-language novel written by a single author, so we exclude works that have

[1] https://en.wikipedia.org/w/index.php?title=Category:19th-century_English_novelists

[2] https://www.gutenberg.org/cache/epub/feeds/

been translated by these novelists and works with multiple contributors.

Using Goodreads authorial metadata cleaned and made readily accessible by Wan et al. (2018, 2019)[3], we discover that 78 of these novelists, who range from classic to obscure, are present on this popular website for book recommendations and reviews. For each author, we take note of their average rating, which indicates readers' perception of their works' quality, and the number of reviews and ratings across all their publications, which indicates their popularity. Unlike Porter (2018), who uses only the number of ratings on Goodreads to measure an author's popularity, we combine both the counts of ratings and reviews because leaving a review indicates a higher degree of engagement with an author's text than simply submitting a rating. To measure the literary prestige of each writer, we follow Porter (2018) in counting the number of academic articles that tag the writer as a primary subject author in the Modern Language Association (MLA) International Bibliography [4]. We then store these three metrics for reception, i.e., quality, popularity, and scholarly acclaim, for the novelists in a single spreadsheet.

To retrieve the 1288 relevant works by these 78 novelists in Gutenberg, we use Angelescu's (2023) Python library to download texts by their Gutenberg identification numbers. Moreover, the library provides a function to remove all sections that are not part of the original novel. As detailed in our methods section, we further clean each text in preparation for our tasks.

4 Lexicon of English clichés

Whereas van Cranenburgh (2018) used a lexicon of over six thousand cliché Dutch phrases for his investigation, we rely on a dataset of over four thousand English clichés compiled in a single plain text file by Reilly (2022) from two sources: the hundreds of expressions curated by Hayden (1999) and the ones assembled by Lepki (2020). Moreover, we expanded it with 19th-century idioms[5] and further edited the file to remove extraneous parenthetical material.

[3] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home?authuser=0

[4] https://quicksearch.library.duke.edu/?utf8=âœŞ&q=MLA+International+Bibliography. Access to the bibliography is provided through Duke Libraries.

[5] https://en.wiktionary.org/wiki/Appendix:English_19th_Century_idioms

5 Methodology

Our approach of replicating part of van Cranenburgh's paper is twofold, namely, data preprocessing and cliché counting.

As previously mentioned, we start off the process of matching the main corpus with the predefined set of clichés (Reilly, 2022) by MWE identification (Kulkarni and Finlayson, 2011; Constant et al., 2017). Since the intended cliché lexicon is well-prepared as a format of one cliché/sentence per line with space-separated tokens, we only need to tokenize the 19th-century novels corpus in a matching format by removing line breaks, segmenting lines per sentence, standardizing punctuations and letters (to lowercase), and replacing hyphens and dashes with a single space. Moving forward, instead of an exact replication of van Cranenburgh's code, our process figures a more adaptable method that allows convenient implementation of Python libraries, such as WordNetLemmatizer[6] and pos_tag[7] by Natural Language Toolkit (NLTK). Specifically, we manage stemming and lemmatization in both the novels and the clichés data with regular expressions (regex). Several emphases are acknowledged based on the nature of regex translation, for example, removing all alternatives of pronouns, conjunctions, single possessives (" 's "), and determinants, recovering contractions (e.g., "isn't" to "is not"), and reversing negations. Exceptions that fall outside of the categories are reviewed and edited manually based on the specifications in the paper (van Cranenburgh, 2018). Additionally, we made the decision to remove clichés that, after preprocessing, were only one word, due to the increased chance of false positives. After both the file of clichés to search for and the files for each author, which contain the text of their novels, were put into the same format, as just described, we were able to search through these files and count the number of occurrences of each cliché for each author.

Essentially, van Cranenburgh adopts a manual/lexicon-based cliché detection strategy in which a list of cliché expressions is required. To count the clichés in our corpus of novelists, we treat each author as one sample unit and summarize the frequency of detected clichés in all of their novels. To contextualize this across authors with

[6] https://www.nltk.org/_modules/nltk/stem/wordnet.html

[7] https://www.nltk.org/api/nltk.tag.pos_tag.html

varying sample sizes, we count the cliché cover, or the percentage of words in their works that are part of a cliché. Additionally, while finding these results, we noted the length of each cliché found and the number of times each cliché was encountered across all texts. Moreover, the original paper visualizes simple linear regressions of the number of clichés with literariness and quality scores (van Cranenburgh, 2018), but we decided to utilize our collection of authors' scholarly and popular reception aiming at a picture of the public domain. Details including post-processing counts can be found in the results section.

6 Results

6.1 Cliché Characteristics

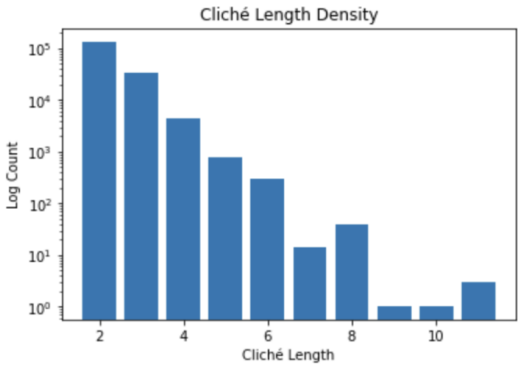

Figure 1: Bar chart comparing the length of clichés versus the log counts of their frequencies

When looking at the most commonly occurring clichés, some identifiable and interesting results include things like, "matter of fact," "flesh [and] blood," "on [the] other hand," "hold tongue," "to tell truth," "for god sake," and "over [and] over again."[8] We found that most of the clichés we found were relatively short (after taking into account the length after cleaning and normalizing), but that there were matches with lengths up to eleven words long (see corresponding figure for length versus log(count of occurrences)). For example, the phrase 'course of true love never did run smooth' had eight matches across all texts for all authors. We were pleased to see that expressions of varying lengths were found throughout all authors' texts after making preprocessing decisions for both the texts and cliché expressions.

[8]See the top 250 clichés found here: https://github.com/amycweng/clichÃĺs-and-19th-Century-English-Novelists/blob/main/clichÃĺ_counts_collection.ipynb

6.2 Cliché Density and Reception

We evaluated the correlation between the log of each author's cliché ratio (total number of cliché count of each author divided by their number of words) and his or her reception, separated into quality, popularity, and scholarly acclaim, with a simple linear regression. The response variables we are interested in are the log of the combined review and rating count, log of the number of corresponding MLA entries, and average rating, and the predictor variable is cliché ratio.

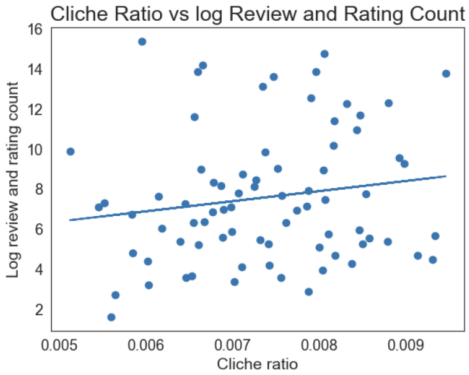

Figure 2: Scatterplot of cliché ratio versus log review and rating count, and line of best fit.

The equation for cliché ratio vs log review and rating count is $3.84 + 505.85 * cliche.ratio$. That is, for every 1 increase in cliché ratio, there is an $e^{505.85}$ increase in the number of reviews and ratings for the author. The linear regression output has a p-value of 0.170 and a residual standard error of 3.847. Using a p-value cutoff at > 0.05, we cannot reject the null hypothesis. The high residual standard error means the linear model is not a good fit on cliché ratio vs log review and rating count.

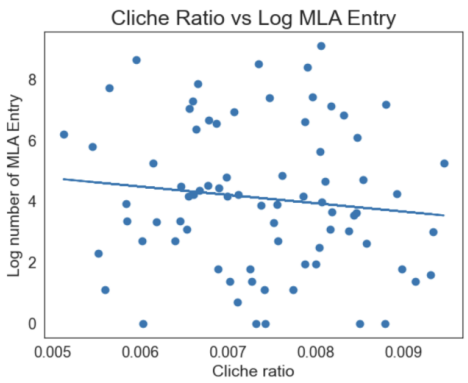

Figure 3: Scatterplot of cliché ratio versus log MLA entry, and line of best fit.

The equation for cliché ratio vs log MLA entry is $6.125 - 273.31 * cliche.ratio$. That is, for every 1 increase in cliché ratio, the number of MLA entries decreases by $e^{273.31}$. The p-value is 0.302 and the residual standard error is 2.364. This is the worst performing model in terms of p-value, which means the linear fit on log MLA and cliché ratio is not good enough to explain their relationship.

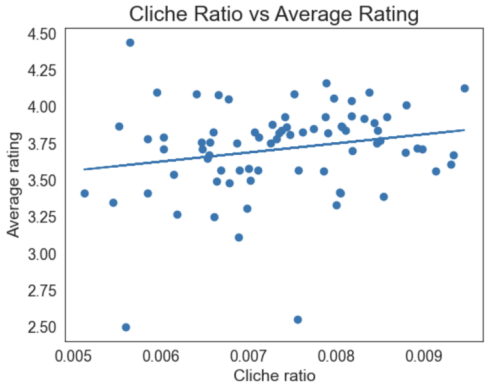

Figure 4: Scatterplot of cliché ratio versus average rating, and line of best fit.

Lastly, the equation for cliché ratio vs. average rating is $3.26 + 61.84 * cliche.ratio$, which means for every 1 increase in cliché ratio, there is an $e^{61.84}$ increase in the average rating of the author. This fit has a p-value of 0.0756 and residual standard error of 0.3088, which is the lowest among the three models. We are most confident that there is a positive linear relationship between average rating and cliché ratio.

We found that the variables of average rating and combined review and rating count are positively correlated with the cliché ratio variable amongst the authors in our dataset, whereas the count of MLA entries is negatively correlated. However, the correlations are weak for the review and rating count and the number of MLA entries. Nevertheless, our results are somewhat in line with what we might predict and is similar to van Cranenburg's findings. The variable predicting "literariness," the number of MLA entries, predictably is correlated with fewer clichés, which makes sense in the context of dislike for clichéd writing in academic and prestigious literary contexts. Unlike van Cranenburgh (2018), we do not find a negative correlation between cliché-ness and quality. Noting that quality and popularity are positively correlated with cliché is an interesting counterpoint to the aforementioned paper and warrants additional study across English novels, particularly throughout different time periods.

6.3 Discussion

However, there are several key limitations in assessing the cliché densities of texts by an author using a manually compiled and then preprocessed lexicon. The lexicon may lack sufficient coverage for a corpus of nineteenth-century novels because many expressions only appear in modern contexts, such as "phone it in," or became clichés only after their first appearance in famous pieces of literature, such as "Nature's first green is gold" from Robert Frost's 1923 poem "Nothing Gold Can Stay." Moreover, after preprocessing, some expressions, unfortunately, become single words, which means we exclude them from our queries to minimize the number of false positives. Some of these expressions are also clichés by virtue of their punctuation—such as "As if!"—but we do not preserve this complexity after stripping out all punctuation except apostrophes from the expressions and texts. After lemmatization and removal of certain parts of speech, phrases may become so general that they yield false positives, such as "it will do" becoming "will do." However, we still feel that preprocessing is necessary to reduce the runtime required to process all texts and capture as many variations in these expressions as possible, and we observe that most expressions still retain their clichéness and specificity after preprocessing. Moreover, we were able to reduce the number of errors by updating our lexicon after running our initial experiments. We removed the clear outliers that get overly high hits due to their vagueness, such as "will do," "to own," and "a if." The latter comes from "As if!" because the lemmatizer removed the "s" from "as."

One well-documented method for addressing the problem of cliché coverage and relevance is to perform an automatic cliché assessment of a text using n-gram counts from a large external corpus (Cook and Hirst, 2013). Indeed, van Cranenburgh (2018) himself replicated Cook and Hirst's methodology with Dutch novels and reference corpus, concluding that higher order high-frequency n-grams from a reference corpus, especially ranging from bigrams to 4-grams, appear more often in known clichéd texts than less clichéd ones. Thus, evaluating the n-gram distribution of texts may help verify or validate the results of a lexicon-based approach. If we were to reproduce the same method,

we would follow Cook and Hirst (2013) in using the Google Books N-Gram Corpus (Michel et al., 2011), specifically counting the high-frequency 2-5-grams which occur in texts published within a time period contemporary to our texts of interest. We would define this time period to be 1745 and 1959, which respectively mark the earliest year of birth and the latest year of death for our set of novelists. Future steps in an expansion of this paper would be to try this n-gram method, which we know is feasible but very time-consuming due to the sheer massiveness of the N-Gram Corpus. However, the n-grams approach will not reveal anything about the characteristics of clichés found within the texts because not all high-frequency n-grams are themselves clichés.

7 Conclusion

Although we began our investigation intending to perform a replication study of van Cranenburgh's 2018 paper on clichés, we made many modifications due to the nature of the data accessible to us in English. An ideal replication study would be with contemporary novels separated into clear genres, but we can only work with data in the public domain. Moreover, we cannot directly investigate the intuitive assumption that more literary novels use less clichéd and more original language because we have no access to ratings of literariness as van Cranenburgh did. Instead, we investigated whether well-received novelists overall use more original language than the more obscure and poorly-rated ones, finding a strong positive correlation between cliché ratio and average rating and weaker correlations for scholarly acclaim (negative) and the combined rating and review count (positive). Overall, these findings must be taken into context with their language of origin and time of writing. We recommend future research into other periods of English novels and perhaps among books in other languages.

References

Radu Angelescu. 2017. *GutenbergPy [Python]*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook and Graeme Hirst. 2013. Automatically assessing whether a text is cliched, with applications to literary analysis. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 52–57, Atlanta, Georgia, USA. Association for Computational Linguistics.

Laura Hayden. 1999. *Clichés: Avoid Them Like the Plague*.

Lisa Lepki. 2020. *The Internet's Best List of Clichés*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

J. D. Porter. 2018. *Popularity/Prestige: A New Canon. Stanford Literary Lab*.

Lee Reilly. 2022. *List of clichés (English)*.

Alex G. Smith, Christopher X. S. Zee, and Alexandra L. Uitdenbogerd. 2012. In your eyes: Identifying clichés in song lyrics. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 88–96, Dunedin, New Zealand.

Nir Sweed and Dafna Shahaf. 2021. Catchphrase: Automatic detection of cultural references. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7, Online. Association for Computational Linguistics.

Andreas van Cranenburgh. 2018. Cliche expressions in literary and genre novels. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 34–43, Santa Fe, New Mexico. Association for Computational Linguistics.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.