Neologisms as a Method of Measuring Lexical Innovation

Niek Holter S3008363
Length: 3-4 pages

Abstract

This paper investigates whether Science Fiction texts exhibit a higher frequency of neologisms compared to Romance literature. Employing exclusion dictionary architecture (EDA) alongside established natural language processing (NLP) techniques provided by SpaCy, we systematically identify and validate lexical neologisms within texts from the early 20th century. By analyzing selected novels from both genres, we anticipate that science fiction authors utilize neologisms more frequently, reflecting genre-specific needs to articulate novel concepts and futuristic contexts. Our methodological approach includes detailed data preprocessing, candidate neologism identification, and manual validation steps through Google's n-gram viewer. Results indicate a substantially higher frequency of neologisms in science fiction (0.44%) compared to romance novels (0.085%), supporting the hypothesis that lexical creativity is more prevalent in science fiction.

1 Introduction

Neologisms play a critical role in linguistic evolution, particularly as society and technology evolve and new concepts emerge. Literary genres differ significantly in their tendency to adopt and create new lexical items. Science fiction literature, especially, is inclined to coin new terms to denote futuristic technologies, alien species, or imaginative concepts. Romance literature, on the other hand, is more grounded in reality and focuses more on interpersonal relationships in familiar settings, potentially relying less on invented terminology. The contrast between these two genres offers us an opportunity to investigate the differences in lexical creativity between genres.

It is important to note that the very notion of "neology" can be "fuzzy and hard to define", with different linguists applying varying criteria (Jamet and Terry, 2018). For instance, Jamet and Terry (2018) mention that some would consider a given occurrence a neologism, some a nonce-formation or hapax. Furthermore, Pruvost and Sablayrolles (2003) make the distinction between lexical neology (a new signifier for an existing or non-existing meaning) and semantical neology (a new meaning for an existing signifier). Since we are interested in measuring lexical creativity and innovation across different genres, we focus on lexical neologisms and consider nonce words, typically viewed as temporary or one-time formations, as neologisms as well, since they still introduce newly formed expressions within a text.

The central research question guiding this investigation is:

Do texts authored by science fiction writers contain a higher frequency of neologisms compared to those authored within the romance genre?

Our hypothesis states that:

The frequency of neologisms will be significantly higher in science fiction texts than in romance literature.

This is supported by (Poix, 2018), who argue that ”lexical innovation is essential to the fantasy-driven world of children's books," as science fiction novels also incorporate this fantasy-driven element. We aim to provide evidence in support of this hypothesis through a combination of established architecture (EDA), NLP applications, and manual evaluation.

2 Related Work

Automatic detection of neologisms commonly relies on an *exclusion dictionary architecture* (EDA), which filters known words using historical dictionaries or lexica so that remaining unknown words are flagged as potential neologisms (Cartier, 2017). EDA is effective for identifying lexical neologisms but has limitations, such as mistakenly identifying proper nouns, spelling mistakes, or corpus artifacts as novel words (Cartier, 2017). To mitigate these issues, previous work commonly integrates additional filters, such as Part-of-Speech (POS) tagging and Named Entity Recognition (NER), to further refine candidate lists. For example, (Zalmout et al., 2019) combine frequency heuristics with exclusion lists to achieve high-precision neologism detection , and explicitly remove named entities using spaCy's NER tool . They emphasize that neologisms are typically absent from "traditional dictionaries or language lexica", necessitating such external references to identify truly novel words.

3 Data

To analyze neologism usage, we collected the texts of six novels: three science fiction and three romance novels, all written in the early 20th century (1907–1912). Focusing on a narrow publication period controls for historical linguistic variation, allowing a fair genre comparison. Each book is by a different author to reduce author-specific style bias, and all were regarded as influential or notable in their genre. Table 1 lists the selected novels and their metadata. We obtained the texts from Project Gutenberg, using the "Gutenberg-dammit" repository, which provides pre-processed Gutenberg texts with most boilerplate removed and useful metadata filters. This facilitated genre-specific retrieval of novels and ensured cleaner input texts.

Pre-processing Each text was cleaned to remove residual Gutenberg metadata (e.g., prefaces, tables of contents, and chapter headers). We then processed the novels with the spaCy NLP toolkit (using the high-accuracy en_core_web_trf model) to tokenize the text and enrich each token with linguistic metadata: lemma (base form), part-of-speech tag, and named entity label. This produced, for each novel, a structured list of tokens with annotations. From these, we built a vocabulary list for each genre containing all unique lowercased tokens. We filtered out tokens identified as punctuation, numbers, stop words, and those tagged as proper nouns (PROPN) or otherwise labeled as named entities by spaCy. In total, the romance novels yielded 14,181 unique filtered tokens, and the science fiction novels yielded 13,502 unique tokens. These vocabulary sizes will be used for normalizing the neologism frequency later on.

To filter out non-neologisms from these vocabularies we utilize EDA based on previous work we referenced earlier. To this end we compiled two exclusion dictionaries. The first is a digitized 1901 Chambers Twentieth Century Dictionary, from which we extracted approximately 42,000 unique headwords. The second is a corpus-derived lexicon: we aggregated all words from a large collection of 19th-century English texts (specifically, all Project Gutenberg works by authors who died before 1900). This resulted in a list of about 1.6 million unique word types that were in use before the 20th century. The rationale is that any word present in either dictionary is likely not a neologism by the time of our novels (1907–1912).

Neologism detection Now to obtain our list of candidate neologisms we simply intersect each genre's token list with the exclusion dictionaries and filter out any token that appears in either reference list. This results in 63 candidate Romance neologisms and 187 candidate Science-Fiction neologisms. The manageable number of candidates and our project's scope allowed us to manually verify each token. Each candidate word was checked against Google's Ngram Viewer and other sources for evidence of prior usage. We considered a candidate a "true neologism" of the period if it showed negligible usage before 1900 and only rose to prominence afterward or was not indexed by Google's Ngram Viewer without being a typo (often times nonce-formations).

Table 1 provides a summary of the data used in this study.

Genre	Title	Author	Year
Science Fiction	*The War in the Air*	H. G. Wells	1908
Science Fiction	*A Princess of Mars*	Edgar Rice Burroughs	1912
Science Fiction	*The Night Land*	William Hope Hodgson	1912
Romance	*Three Weeks*	Elinor Glyn	1907
Romance	*The Shuttle*	Frances Hodgson Burnett	1907
Romance	*The Rosary*	Florence L. Barclay	1909

Table 1: Overview of novels selected for analysis, categorized by genre, author, and year of publication.

All data processing and analysis code used for this project is available in our public repository (GitHub)[1] for reproducibility.

4 Results and Analysis

In our hypothesis we predicted a clear difference in neologism usage between the two genres in favor of Science Fiction texts. After applying the neologism detection pipeline and manual validation, we are left with 59 distinct neologisms in the science fiction texts, compared to about 12 neologisms in the romance texts. To account for the slightly different vocabulary sizes, we normalize these counts as a percentage of unique tokens per genre: approximately 0.44% of the unique word types in the SF novels are neologisms, versus only about 0.085% in the romance novels.. See Table 2 below for an overview of these results.

Genre	Candidate neologisms	Validated neologisms	Normalized neologisms
Science Fiction	187	59	0.44%
Romance	63	12	0.085%

Table 2: Overview of neologism counts and normalized frequencies by genre.

Discussion The results confirm our hypothesis that science fiction authors create significantly more neologisms than romance authors, highlighting genre as a key factor influencing lexical creativity. science fiction authors introduced 59 new words in our sample, compared to just 12 in Romance. This disparity is likely due to science fiction's need for detailed world-building, authors frequently invent terminology to depict futuristic technology, alien societies, or unique concepts, making these worlds believable and distinct from reality. Examples include ”jeddak” (a leader title from Edgar Rice Burroughs' Mars novels) and ”drachenflieger” (an aviation-related term used by H. G. Wells), both illustrating how science fiction neologisms can gain explanatory significance within narratives.

Conversely, romance novels typically explore more grounded, realistic settings emphasizing personal relationships, which reduces the need for new vocabulary. Their few neologisms, such as ”waldflute” or ”blimme,” tend to be playful, stylistic inventions without significant explanatory roles or lasting linguistic impact. Thus, romance remains largely within established vocabulary boundaries.

The genre-specific usage of neologisms demonstrates how literary themes and settings influence lexical innovation. science fiction tends to introduce terms that have potential for wider adoption if their underlying concepts become culturally relevant, similar to how the term ”robot” was invented by Karel Čapek in his 1920 play R.U.R. (Rossum's Universal Robots). In contrast romance's creativity primarily emerges through narrative style and emotional depth rather than new lexical items.

5 Conclusion

This study set out to explore whether science fiction authors use more neologisms than romance authors, and our findings strongly suggest that they do. By applying an exclusion-based neologism detection method (EDA) and examining a genre-representative corpus of early 20th-century novels, we found that science fiction texts had a markedly higher frequency of lexical innovations (neologisms). The science fiction authors in our sample coined numerous terms to describe their novel inventions and worlds, whereas the romance authors introduced very few new words. These results support the hypothesis that genre influences lexical innovation. Specifically, imaginative genres like science fiction serve as hotbeds for neologism creation, while grounded genres like romance tend to stick to established vocabulary.

Our approach combined computational filtering with manual verification, yielding high-confidence results on a small scale. One key contribution of this work is demonstrating a methodology to compare neologism frequencies across genres. We showed that leveraging historical dictionaries and NLP tools (like POS tagging and NER) is effective in isolating true neologisms (Zalmout et al., 2019), even in literary texts, confirming observations from prior studies that such exclusion lists can recover new words with high precision.

The small sample size in this study means our numerical estimates are illustrative rather than definitive. The observed neologism rates (0.44% for science fiction vs. 0.085% for romance) reflect the texts we analyzed, and results could vary with a larger or different set of texts. A broader corpus, including more authors and genres, could clarify whether the identified trend is consistent across lit-

[1] `https://github.com/muniekstache/intro_research_resit.git`

erature. Additionally, our focus was only on lexical neologisms; exploring semantic innovations (new meanings given to existing words) could further illuminate how creativity manifests differently across genres. Finally, Examining other genres, like fantasy literature, may also provide insights into whether imaginative literary genres consistently produce more new vocabulary compared to genres like romance.

References

Cartier, E. (2017). Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 95–98.

Jamet, D. and A. Terry (Eds.) (2018). *Lexical and Semantic Neology in English*, Volume 12. Lexis.

Poix, C. (2018). Neology in children's literature: A typology of occasionalisms. *Lexis 12*.

Pruvost, J. and J. Sablayrolles (2003). Les néologismes, n 3674. *Paris,Que sais-je*, 17.

Zalmout, N., K. Thadani, and A. Pappu (2019). Unsupervised neologism normalization using embedding space mapping. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 425–430.