2013

Comparative Stylistic Fanfiction Analysis: Popular and Unpopular Fics across Eleven Fandoms

Victoria L. Rubin
Western University, vrubin@uwo.ca

Vanessa Girouard
Western University

Girouard, V., Rubin, V.L. (2013). Comparative Stylistic Fanfiction Analysis: Popular and Unpopular Fics across Eleven Fandoms
In the Proceedings of the Annual Conference of the Canadian Association for Information Science, 6-8 June, Victoria, BC

Comparative Stylistic Fanfiction Analysis: Popular and Unpopular Fics across Eleven Fandoms

Vanessa Girouard, Victoria L. Rubin
Language and Information Technology Research Lab (LIT.RL)
Faculty of Information and Media Studies
University of Western Ontario
North Campus Building, Room 260,
London, Ontario, Canada N6A 5B7Affiliation
vgirouar@gmail.com , vrubin@uwo.ca

Abstract: This study analyses 545 sample fanfiction stories (*fics*) in their stylistic feature variation by popularity and across eleven 'fandoms' in creative writing forums. Lexical richness, average sentence and paragraph lengths are isolated as promising measures for a text classifier to use in predicting a fic's likely popularity in its fandom.

Résumé: Cette étude analyse un échantillon de 545 chapitres d'œuvres de fanfiction (fics) selon leur variation stylistique et leur popularité dans onze 'fandoms' différents. La richesse lexicale, longueur moyenne de phrase et longueur moyenne de paragraphe ont été choisis comme traits stylistiques propres à différencier les fics populaires des fics impopulaires.

1. Introduction

Library Science has traditionally sought to understand the motivations behind readers' choices, while Information Science is the enterprise of trying to make sense of the overwhelming amounts of information available to users. This study combines both efforts in studying the less documented role of stylistics in story popularity and by attempting to devise a tool that would help readers narrow down their choices when it comes to choose which work of cyber-literature (a vast and growing corpus) they want to read next.

Our corpus is fanfiction, the transformative practice of writing fiction based on elements (characters, settings, etc.) of an existing source text, i.e. any professional material such as a movie, book, video game, etc. (Wenz, 2010; Pimenova, 2009). While the popularity of

a novel is difficult to quantify, online fanfiction provides a convenient corpus: fanfiction stories (fics) are labelled for popularity by readers via 'favorites'. This facilitates stylistic analysis using Natural Language Processing (NLP) tools and makes fanfiction the ideal dataset for a supervised text categorization (or document classification) task.

We claim that: (i) popular and unpopular fics differ stylistically and that (ii) isolated linguistic features of popular and unpopular fics can inform a text classifier (a system that would take any new amateur text of fiction as input and predict its likely popularity). Features selected for this pilot are: (i) lexical richness, (ii) average sentence length and (iii) average paragraph length.

2. Methodology

Our dataset consists of the first chapters of 545 fics available on Fanfiction.Net, the largest online fanfiction archive, narrowed down by four criteria (Table 1). Analysis was conducted per source text; each set is referred to as source text fic set (STFS), the details of which are detailed in Table 2.

Criteria	**Value**	**Justification**
Total length of fic	> 60,000 words	Length of the average paperback / light read
Language	English	
Source text popularity	Source text must have > 50,000 works archived on Fanfiction.net	Allows for larger, statistically sound sample
Current Status	Story: Complete	Incomplete stories are less likely to be read / favored

Table 1. Fanfiction Subset Considered for Data Sample

Source Text Type	**Fandom name**	**Number of *Most* Popular Fics Sampled**	**Number of *Least* Popular Fics Sampled**
Book	*Harry Potter*	**25** top fics per each fandom, first chapter*	**25** (out of the 50 top) fics per each fandom, first chapter
	Twilight		
Anime/Manga	*Naruto*		
	Hetalia-Axis-Powers		
	Bleach		
	Yu-Gi-Oh		
	Inuyasha		
TV	*Glee*		
	Supernatural		
Games	*Kingdom Hearts*		
	Pokemon		

Table 2. Data Sample Characteristics.

**Exceptionally, Harry Potter showed 20 rather than 25 results on the first page.*

Popularity is measured in favorites count. 'Favorites' indicate the amount of registered Fanfiction.Net users listing the fic in their publicly viewable 'Favorite Stories' list.

Lexical richness, sentence length and paragraph length were selected as likely stylistic factors that fan readers take into account when passing judgement on fics, according to fan readers' opinions posted on Fanfiction.Net forum threads.

Lexical richness is the ratio of total word count divided by the number of unique words in a document, with values ranging from 0 to 1. Unusually small / large word inventories may entail word repetition; high occurrence of jargon, synonyms, and spelling errors, which can detract the reader from enjoying the piece.

Sentence and paragraph length are measured in characters. Both can affect reading comprehension, making the text more or less difficult to parse for the reader.

Each fic was processed algorithmically to extract story text and strip HTML tags, then tokenized into meaningful linguistic units using existing and custom-built tools. Word and sentence units were tokenized using Bird, Klein, and Loper (2009)'s Natural Language ToolKit (NLTK). Paragraphs were tokenized using Girouard's own algorithm.

3. Results

Data distributions were illustrated with histograms and scatterplots with linear regressions. Outliers mostly belonged to unpopular fic values. Popular fic values are represented in green; unpopular fic valuesare represented in red; histogram overlay is brown.

Lexical richness of popular fics falls within a narrower range of values than that of unpopular fics. The curve for unpopular fics tilts leftwards, while the curve for popular fics tilts rightwards, for seven STFS, hinting that popular fics have a larger word inventory and unpopular fics show more word repetition.

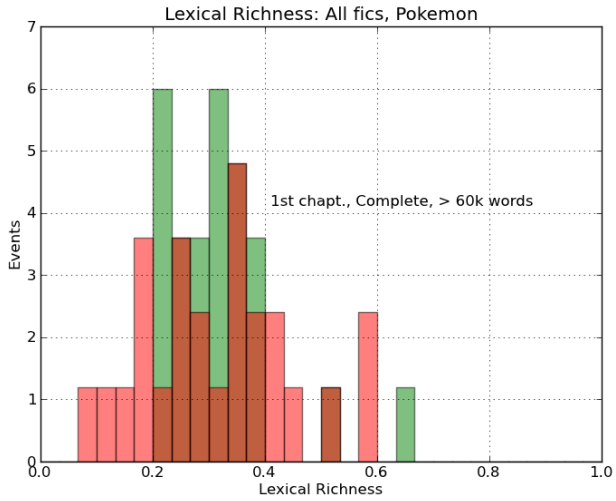

Figure 2. Lexical Richness of Popular and Unpopular Fics for Source Text: *Pokemon.*

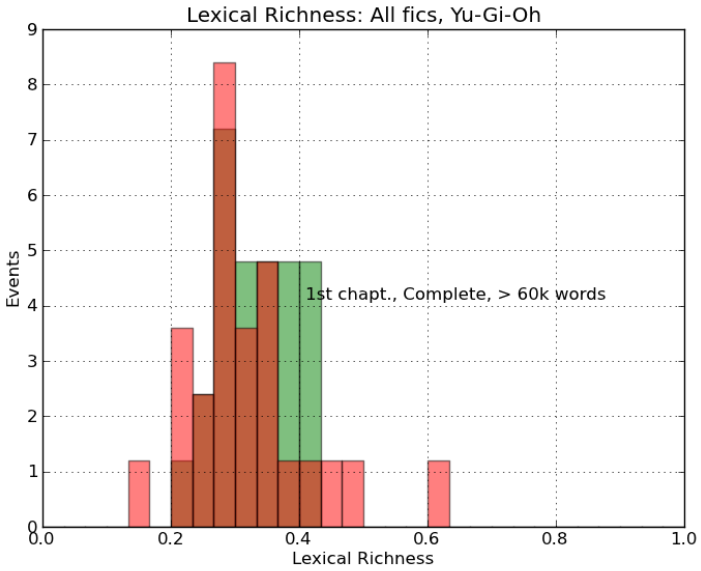

Figure 3. Lexical Richness of Popular and Unpopular Fics for Source Text: *Yu-Gi-Oh.*

Lower values for average sentence length generally map to unpopular fics. Higher values generally map to popular fics. Sentence length is the only feature where popular fic values sometimes strand from the cluster (larger values). For some STFS, unpopular fics have larger paragraphs and shorter sentences than their popular counterparts. Popular fics map closer to their regression line (which is similar across STFS) than unpopular fics.

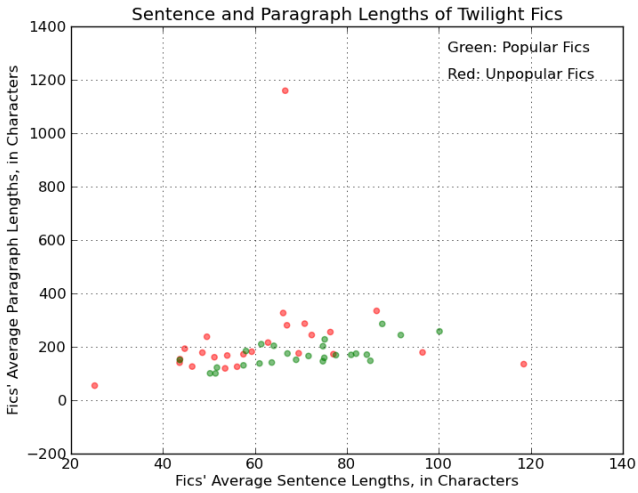

Figure 4. Average Sentence and Paragraph Lengths for *Twilight* Fics, Topmost Outlier Included

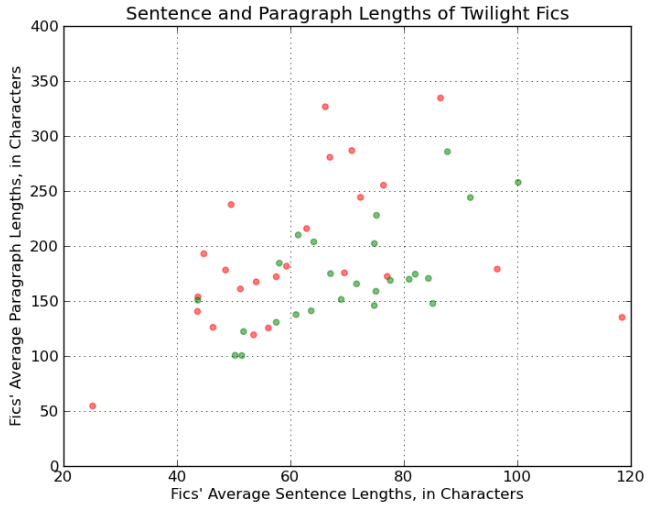

Figure 5. Average Sentence and Paragraph Lengths for *Twilight* Fics, Topmost Outlier Excluded

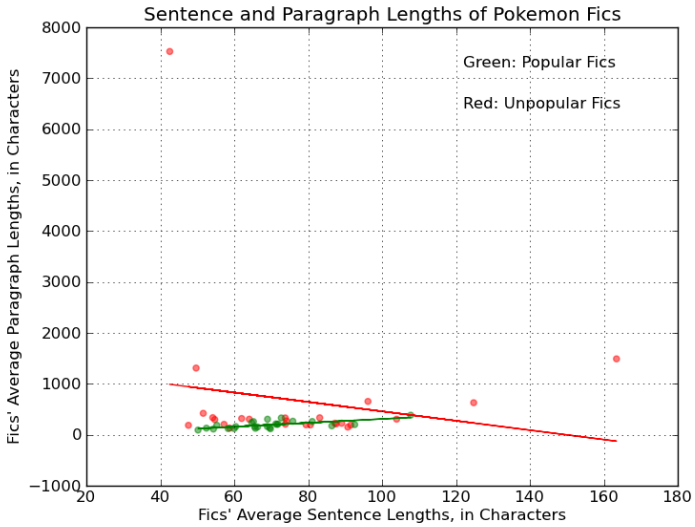

Figure 6. Average Sentence and Paragraph Lengths for *Pokemon* Fics, with Regression Line and Outliers.

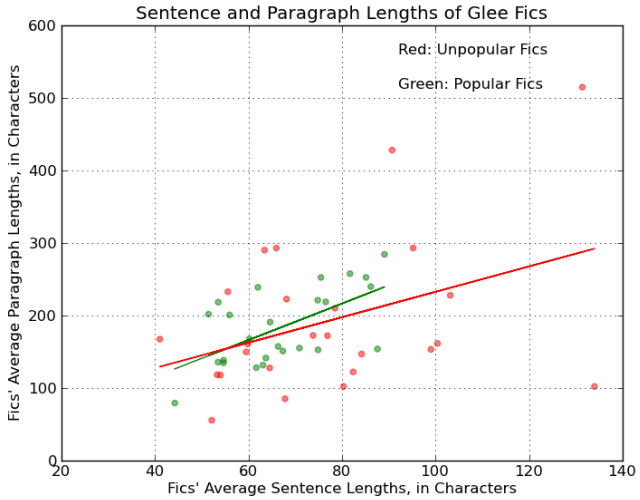

Figure 7. Average Sentence and Paragraph Lengths for *Glee Fics*, with Regression Line and Outliers

4. Conclusions, Limitations, Future Work

In addition to assisting readers in their reading choices, a text classifier that assigns the categories ‘likely to be popular’ and ‘likely to be unpopular’ could be, for authors aspiring to have their works of fiction published, a preliminary step to soliciting professional agents, editors and publishers.

We showed preliminary evidence that stylistics correlate with popularity for online fanfiction corpora. Unpopular fics tend to be more stylistically diverse than popular fics in lexical richness as well as average sentence and paragraph length.

Our preliminary findings, subject to further verification and statistical testing, are promising for developing automatic capability to predict fanfiction popularity. The novelty in Information Science lies in applying NLP methods to a rarely studied corpus: fanfiction.

References

Bird, Steven, Ewan Klein and Edward Loper. 2009. Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit. Beijing: O'Reilly Media.

Pimenova, Daria. 2009. Fanfiction: Between text, conversation, and game. In Internet Fictions, Hotz-Davies, Ingrid, Anton irchhofer, and irpa Lepp nen, eds. Newcastle upon Tyne: Cambridge Scholars Pub.

Wenz, Karin. 2010. Storytelling goes on after the credits: Fanfiction as a Case Study of Cyberliterature. In Reading moving letters: Digital literature in research and teaching : A handbook. i anowski, Roberto, rgen ch fer, and Peter Gendolla, eds. Vol. 40. Bielefeld: Transcript Verlag.