Fan Fiction Stories on Male Homosexuality: Women's Silent Revenge

Siyi Wu*
UC Berkeley
siyiwu@berkeley.edu

Violet Yao*
UC Berkeley
violetyao@berkeley.edu

1 Introduction

According to the statistics in 2010 from FanFiction.net, the world's largest fan fiction archive and forum, around 78% of fan fiction writers are female. The largely female composition of media fandom shows that media fan fiction has been a female undertaking. Moreover, as fan fiction is created and read anonymously by anyone who has passion, it is large in the number of works and diverse in languages involved. This makes fan fiction a great source for us to observe and analyze how females across cultures imagine genders differently or similarly. Writing anonymously and freely, females are empowered to rework on popular texts and to reflect their own interests against the reality with patriarchal order and masculine dominance. By analyzing fan fiction stories, we see how female perspectives are put into source texts which are generally written by and about men, challenging norms of gender and sexuality. In our project, we will sample 9450 fan fiction works in Chinese and English from 5 fandoms. We will extract names of characters using named entity recognition and coreference resolution, utilize social network analysis (SNA) to track character relationships, and use the topic modeling to characterize ways of expressing characters in fan fiction stories.

2 Literature Review

Henry Jenkins' seminal research around 30 years ago on fan culture drew the attention from academia on the topic of fan studies. In *Textual Poachers* (Jenkins, 1992), Jenkins illustrates readers as "nomads" who are always on the move, and "rather constantly advancing upon another text, appropriating new materials, making new meanings." He argues for the originality and creativity hidden inside fan fiction which usually "make inter-textual connections across a broad range of media texts."

Fan fiction' value of investigation not merely originates from their literary exploration, imaginative expansion, and creative interpretation, but also derives from their unique female identity. It is clear from statistics that fan fiction is largely penned by female authors, which is not difficult to understand by putting it back to a larger sociological context. As indicated in *Edging Women Out* (Tuchman and Fortin, 1989) that "before 1840 at least half of all novelists were women; by 1917 most high-culture novelists were men", the trend of declining female authorship reflects the narrowing room for female writers to participate in the field of serious literature. Furthermore, the male domination in popular culture industries makes so many media products such as movies, TV shows, and books center on the heterosexual, masculine gaze. As Laura Mulvey writes in *Visual Pleasure and Narrative Cinema* (Mulvey, 1975), women are characterized by their "to-be-looked-at-ness" in cinema as "spectacle", while man is "the bearer of the look". Under these social circumstances, fan fiction becomes a wonderland for women and queer fans but fail to reflect their own image in their beloved stories.

Scholars have already established many works on m/m slash (a genre of fan fiction that focuses on romantic or sexual relationships between male fictional characters). *Enterprising Women: Television Fandom and the Creation of Popular Myth* (Bacon-Smith, 1991) concerns slash writing that the male characters are actually surrogate women. *Future Men* (Penley, 2014) argues that female m/m slash writers "spontaneously" get the idea of writing their sexual and social utopias through a futuristic and technologized version of the Sacred Marriage of males. Milli and Bamman (2016) also have employed natural language processing (NLP) methods to investigate differences in the characters emphasized and the treatment of gender in fan fiction stories compared to the original works. In this

project, we will computationally analyze popular romantic relationships and characters in fan fiction stories beyond English-written works.

3 Data Collection

Our data is collected in October 2020 from Archive of Our Own (AO3) (https://archiveofourown.org/), one of the internet's most well-known fan fiction archives. In our project, we use fan fiction terms consistent with Milli and Bamman (2016), referring to CANON as the original work on which a fan fiction story is based and STORY as a single fan-authored story for some canon. We select five canons, including *Harry Potter*, *Marvel*, *Sherlock Holmes*, *Les Misérables* and *Módào Zǔshī*.

We deliberately choose the five canons among the most popular fandoms on AO3. Originated in different eras and cultures, these canons focus on a wide variety of themes. While the other four canons are generally well-known to the general public, *Módào Zǔshī* (literal translation: "Grandmaster of Demonic Cultivation") is a more recent work but ranked among top-35 of all AO3 works in 2020. *Módào Zǔshī* is a Chinese novel that depicts a fictional Xianxia world where humans attempt to attain immortality. While selecting canons, We also strive to achieve a balance between works created in western & eastern language cultures, as well as a balance on genre (popular literature & classics). A summary of our data is presented in Table 1. We collect 1000 English-written stories with top hits and 1000 Chinese-written stories with top hits from AO3 for each canon. Note that for *Les Misérables* stories, we only collect 450 works in Chinese since they are the only ones available.

We use an AO3 web scraper developed by Jingyi Li & Sarah Sterman at UC Berkeley.

https://github.com/radiolarian/AO3Scraper

The existing scraper does not support scraping in languages other than English, thus we modify the script a bit to make it work.

Canon	Language	Number of Stories
Harry Potter	English	1,000
Harry Potter	Chinese	1,000
Marvel	English	1,000
Marvel	Chinese	1,000
Sherlock Holmes	English	1,000
Sherlock Holmes	Chinese	1,000
Les Misérables	English	1,000
Les Misérables	Chinese	450
Módào Zǔshī	English	1,000
Módào Zǔshī	Chinese	1,000

Table 1: Statistics of our AO3 dataset.

4 Exploratory Data Analysis

Aside from title and body, we also collect the following metadata for each story: RATING ("Explicit", "Mature", "Teen And Up Audiences", "General Audiences"), CATEGORY (type(s) of romantic relationship in the work, e.g. "Female & Female", "Female & Male", "Male & Male", etc.), RELATIONSHIP (character pairs of the work, e.g. "Harry Potter/Draco Malfoy"), CHARACTER (characters involved in the work), PUBLICATION DATE, STATUS ("Completed", "Updated"), WORD COUNT, CHAPTER COUNT, COMMENT COUNT, KUDO COUNT, NUMBER OF BOOKMARKS and NUMBER OF HITS. Note that all of them are tagged by authors themselves, as a measure to highlight major elements in their works as well as better attract attention from readers. In this analysis, we primarily focus on CATEGORY, RATING, and RELATIONSHIP.

Firstly, we plot each canon's distribution of categories of stories in both languages. We observe that the largest proportion of stories are tagged as "Male/Male" for all five canons and for both English and Chinese. To save space, we only display the graph for *Harry Potter* with the greatest difference between different languages here. In Figure 1, among all *Harry Potter* stories, 78% of 1000 stories in English are tagged as "Male/Male" and 89% of 1000 stories in Chinese are tagged as "Male/Male". We notice that the m/m slash percentage of Chinese stories is higher than that of English, which is also the case for the rest of the stories except for *Les Misérables* with almost equal percentages. This result resonates with the finding from our literature review section that male-male romantic relationship seems to be the biggest interest for fan fiction writers.

Secondly, we explore the relationships between RATING and CATEGORY by languages in Figure 2 and Figure 3. They prove the finding from the literature review section that generally the greatest proportion of Male/Male stories in both Chinese and English are tagged as "Explicit" or "Mature", which means there are sexual descriptions or other

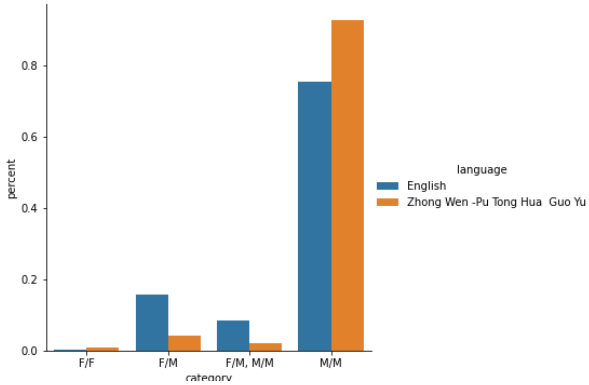

Figure 1: Differences in category popularity between English and Chinese stories for *Harry Potter*

adult contents within the works. One noteworthy point is that for *Les Misérables*, most of the Chinese works are tagged as "General Audience" while English works maintain the general characteristics. One reason for it is the relatively small fandom of *Les Misérables* with only 450 Chinese stories. As a serious literature work written in 19-century France, it has fewer readers and fans than other genres in China. Moreover, by skimming works, we also notice that many stories are not merely about illustrating romantic relationships, but also expressing writers' ideas on political or historical issues. For example, there is a story called *World Ain't Ready* among the top hits which touches realistic social issues such as school bullying, the minority of sexuality, and family problems.

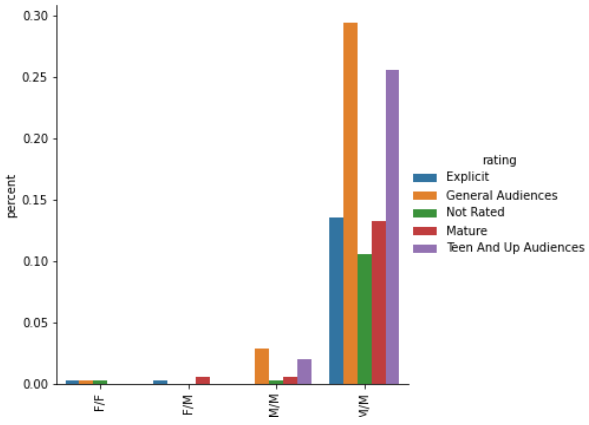

Figure 2: Relationships between RATING and CATEGORY for *Les Misérables* stories written in Chinese

Lastly, we explore the most frequently featured romantic relationships in stories written in both languages. Interestingly, we find a lot of creativity that many popular character pairings emphasized in stories are not established in their original canons.

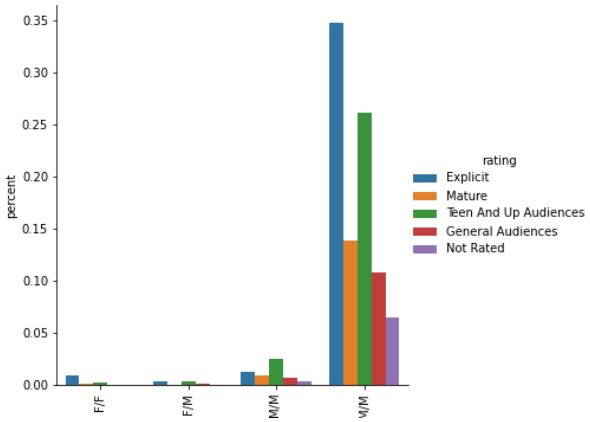

Figure 3: Relationships between RATING and CATEGORY for *Les Misérables* stories written in English

For example, Sherlock Holmes/John Watson (mentioned in around 450 to 500 stories among the total 1,000 stories) is the most popular character pairing in both Chinese and English stories, while it actually does not exist at all in the original *Sherlock Holmes* work. Another interesting point is that writers from different language cultures seem to have different interests in character pairings. As an example, the romantic relationship between Mycroft Holmes and Greg Lestrade is very popular among Chinese writers (mentioned in 100 out of the 1000 stories), but does not receive as much attention from stories written in English (only 11 out of 1,000 stories). We will explore further the different emphasis on character pairings in later sections.

5 Methods

In Labatut and Bost (2019)'s survey paper, the authors stated that building a fictional character network consists of three steps: 1) identification of characters (detection of occurrences and unifying occurrences); 2) identifications of characters' interactions; and 3) construction of the graph. In the methodology section, we follow Labatut and Bost (2019)'s three steps to identify characters by Named Entity Recognition (NER) & Coreference Resolution, identify interactions by counting co-occurrences within a window, and build graphs using NetworkX.

In addition to networks extracted from text, we also utilize RELATIONSHIP metadata (e.g. Harry Potter/Draco Malfoy) tagged by authors. Network analysis based on relationship tags is also presented in the subsection below.

While easily identifying relationships among related characters, we could not find how exactly genders are represented merely from the networks. Muon Nguyen et al. (2013) brought up an interesting idea that we could use topic modeling to better understand discussed content in social networks. Thus, we utilize topic modeling to characterize figures depicted in stories as well as explore on the similarity/differences of language usages of Chinese and English writers. We run topic modeling for the 5 most-mentioned characters in each selected canons, and further analyze lexical diversity by languages and canons.

5.1 Tokenization & Named Entity Recognition

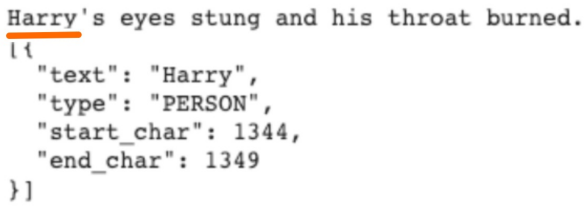

Figure 4: Example of named entity recognition result on a sentence from a *Harry Potter* story in English. Only "PERSON" entity tag is shown.

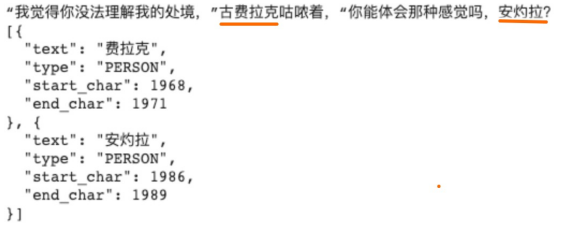

Figure 5: Example of named entity recognition result on a sentence from a *Les Misérables* story in Chinese. Only "PERSON" entity tags are shown. English translation: *"I don't think you understand my situation," Courfeyrac mumbled. "Do you feel the same way I feel, Enjolras¿'* Courfeyrac and Enjolras are extracted.

We utilize Stanza developed by Qi et al. (2020) to tokenize and recognize named entity in stories. Stanza is an NLP pipeline that supports multiple languages. Figure 4 and Figure 5 show two example sentences that we run stanza on.

Results from NER are used to build character dictionaries (5 canons $\times$ 2 languages), where keys are formal English names of all characters and values include lists of possible Chinese/English mentions in stories. We select interest characters based on the most frequent characters from CHARACTER metadata, manually map most-mentioned PERSON entities extracted from NER to interest characters. An example of character dictionary is presented in Figure 6.

```
"""
Example of characters dic
characters_dic = {'Jiang Cheng | Jiang Wanyin': ['江澄','阿澄','晚吟','三毒','江晚'],
 'Meng Yao | Jin Guangyao': ['金光瑶','光瑶','敛芳尊','阿瑶'],
 'Jin Ling': ['金凌', '如兰','孟瑶'],
 'Lan Huan | Lan Xichen | Lan Xi Chen': ['蓝曦','曦','蓝曦臣','曦臣','蓝涣','泽芜君','泽芜'],
 'Lan Si Zhui': ['蓝思追','思追','蓝思'],
 'Lan Wang Ji': ['忘机','蓝湛','含光君','湛','含光'],
 'Nie Huai Sang': ['聂怀桑', '怀桑'],
 'Nie Ming Jue': ['聂明', '聂明玦','明玦','赤锋尊','赤锋'],
 'Song Lan': ['宋岚', '岚','子琼','宋子'],
 'Wei Ying Wei Wuxian': ['魏婴','魏','夷陵','莫玄羽','玄羽','无羡','魏无','魏兄'],
 'Wen Ruo Han': ['温若寒','温若'],
 'Xiao Xing Chen': ['晓星尘','晓星'],
 'Xie Yang': ['薛洋','成美','薛'],
 'Jiang Yanli': ['江厌离','厌离','江厌' ],
 'Wen Ning | Wen Qionglin': ['温宁', '琼林', '鬼将军'],
 'Jiang Fengmian': ['江枫眠', '枫眠','江枫']}
"""
```

Figure 6: Character Dictionary for Módào Zǔshī stories in Chinese.

5.2 Coreference Resolution

Although Stanza does not support Coreference Resolution, we access Stanford CoreNLP from Stanza interface and run Coreference Resolution in English and Chinese. One challenge we encountered is that since our stories are mostly book-length, it takes around 30 minutes to process each story with 20G GPU memory allocated. To ensure efficiency, we decided to randomly sample a continuous block of 5000 tokens from each story, and utilize character dictionaries to unify mentions in coreference chains. Details are described in Algorithm 1.

5.3 Social Network Analysis

We utilize NetworkX (Hagberg et al., 2008) for network construction and Gephi (Bastian et al., 2009) for network visualization. Throughout the previous process, we match each character's possible names mentioned in stories with his/her formal name from the original canon (combined in a character dictionary), as well as a cluster of sentences in which the character appear. We then adapt code from Professor Bamman's lecture on Social Network to compute the number of times two characters co-occur in a window of 25 as edge weights.

https://github.com/dbamman/comphumF20/blob/master/12.networks/ExtractingLiterarySocialNetworks.ipynb

With RELATIONSHIP metadata scraped from AO3, we also construct romantic relationship net-

Algorithm 1: Coreference Resolution

```
initialize charLocationDic ;
for each story in stories do
    ann = client.annotate(story,
      properties=language);
    chains = ann.corefChain ;
    for each chain in chains do
        mentions = chain["mentions"] ;
        mentionedChar = None ;
        for each char in charDic do
            possibleCharacterNames =
              charDic[char] ;
            if any possibleCharacterNames
              appear in mentions then
                mentionedChar = char ;
                break ;
            else
                continue ;
            end
        end
        if mentionedChar exists then
            add (storyId, sentenceId) to char-
              LocationDic[mentionedChar];
        end
    end
end
```

works whose nodes are characters and edge weights are the total number of times a character pairing appear in its canon's RELATIONSHIP metadata.

To avoid confusion, we name network graphs on character pairings as romantic relationship graphs, and network graphs based on direct text extraction as character interaction graphs.

5.4 Topic Modeling

Similar to Network Analysis, this part also makes use of the dictionary of {one character: a list of sentences in which they appear}. We utilize Genism (Řehůřek and Sojka, 2010) to run topic modeling (5 canons $\times$ top 5 characters), keeping top 20 topics each with top 15 tokens.

5.5 Lexical Diversity

$$LexicalDiversity_i = \frac{\sum \text{unique tokens in } Story_i}{\sum \text{tokens in } Story_i}$$

The lexical diversity of a story is measured by the fraction of the number of unique tokens among the total number of tokens. We oversample Chinese Les Misérables stories to 1000 in order to ensure Les Misérables has the same scale as other canons.

6 Analysis

Throughout all network graphs, we generalize several patterns shared by fan fiction stories on all 5 canons. Firstly, each fandom has one or several particularly popular character pairing(s) and almost all of them are male-male relationships. By identifying the thickest and darkest line(s) from romantic relationship graphs, we observe Draco Malfoy/Harry Potter from *Harry Potter*, Grantaire/Enjolras from *Les Misérables*, Loki/Thor from *Marvel*, Sherlock Holmes/John Watson from *Sherlock Holmes*, WeiWuxian-LanWangji from *Módào Zǔshī* and so on. Many of the mentioned are major figures emphasized in the original canons, but there are also minor ones such as Draco, the peripheral character depicted arrogant and spiteful in the original *Harry Potter* series. Fan fiction writers actually pay a lot of attention to minor characters and take efforts to illustrate their relationships. For example, Mycroft Holmes/Greg Lestrade is mentioned a lot in stories. They seem like an unusual pairing because they never appear in a scene together, and thus there is no attraction or sexual tension between them hinted by the original *Sherlock Holmes* series. But it is the lack of any canonical interaction between them that leaves the room for fan fiction writers. They create endless possibilities for them and imagine things like their first meeting, their reason for attraction and development of their romance. Though canons are usually written from a heterosexual perspective, fan fiction writers are free to build ideal parallel universes for their beloved character pairings and re-imagine their lives based on a developed character set.

Secondly, fan fiction writers rarely emphasize female characters in their stories. Hermione is the only female who appears in the most popular character pairings on romantic relationship networks. Further, we notice that her significance diminishes in character interaction graphs, while most of the other characters (all male) within popular character pairings maintain similar positions in both types of network graphs. This fact reveals that aside from her romance, Hermione receives very limited attention from fan fiction writers. In fan fiction stories, female figures are mentioned as supporting characters for male-male romance. Narcissa Black Malfoy, Draco Malfoy's mother from *Harry Potter*,

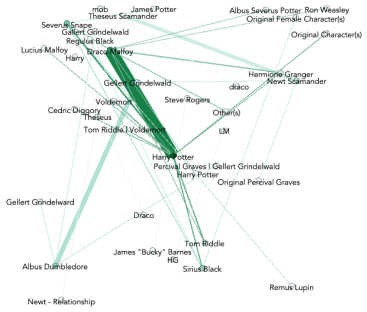

Figure 7: Romantic Relationship Network visualization of *Harry Potter* Stories in Chinese

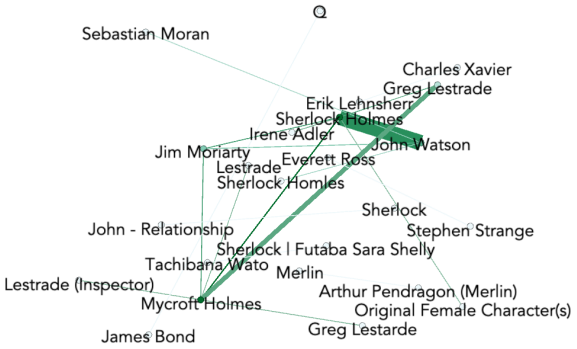

Figure 8: Romantic Relationship Network visualization of *Sherlock Holmes* Stories in Chinese

Figure 9: Romantic Relationship Network visualization of *Módào Zǔshī* Stories in English

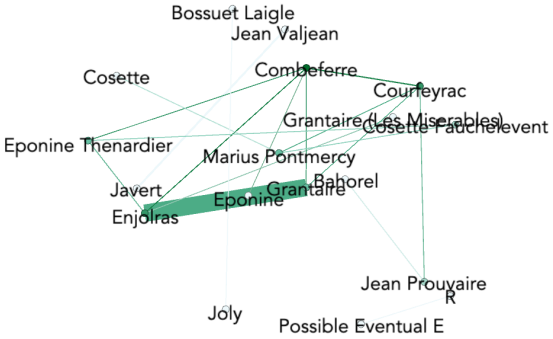

Figure 10: Romantic Relationship Network visualization of *Les Misérables* Stories in English

doesn't appear in the romantic networks but holds a relatively central position in character interaction graphs. This is because she appears in most of Draco Malfoy/Harry Potter stories as the disdainful and haughty mother standing in the middle of her son and his lover Harry. It is a surprising result because we would naturally assume that female fan fiction writers would be interested in establishing female characters who are usually marginalized in original works, while it seems not. Then, what interests them?

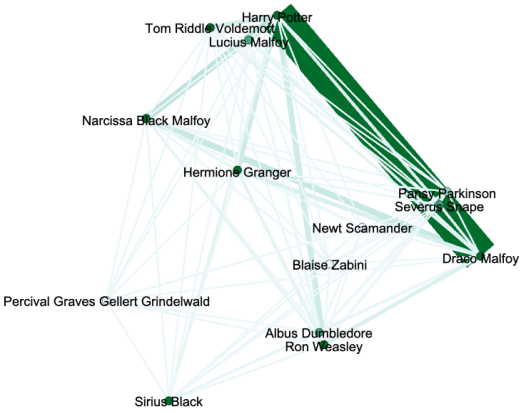

Figure 11: Character Interaction Network visualization of *Harry Potter* Stories in Chinese

Our third finding is that popular character pairings all seem "impossible". Voldemort and Dumbledore each represent the brightest and darkest side which has irreconcilable conflicts in the Harry Potter magic world, but they are the second most popular character pairing in Chinese Harry Potter fan fiction stories. Loki and Thor, set as brothers in

the Marvel universe, attract many fans to write love stories between them. For fan fiction writers, male-male romance between opponents, mentor-mentee, enemies and brothers which would rarely exist in real world naturally comes with dramatic tension and a sense of tragic beauty. Writing about "impossible" and "shameful" romance is the process of helping their beloved characters overcome all kinds of burdens in the reality and achieve an ideal utopia with freedom and love. The community of reading and writing fan fiction stories is thus like a secret safe space where silent revolution led by women on patriarchal order and heterosexual dominance happens.

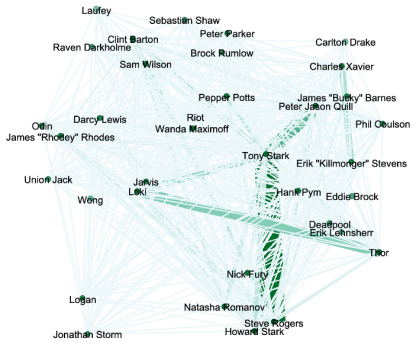

Figure 12: Character Interaction Network visualization of *Marvel* Stories in English

To further analyze how characters are depicted in fan fiction stories, we use topic modeling aiming to characterize each figure. Interestingly, we find that among the corpus created by topic modeling, there are rarely adjectives, but a lot of character names and verbs. For each character, the character name that most frequently appears in his/her corpus is usually the most popular romantic partner of him/her in fan fiction stories. For example, in Grantaire's list of words, Enjolras is the character name that comes up the most, and Grantaire/Enjolras is the most popular character pairing in stories on *Les Misérables*. Verbs mentioned in the corpus are divided into three types: verbs signaling drastic conflicts, implicit language on sex, and daily verbs (such as "say", "look", "feel", etc). In the corpus for Draco Malfoy, there are a lot of occurrences of "Harry", verbs including "kiss", "simper", "love", "curse" in both English and Chinese, as well as words for sex scenes such as "cock""touch""smack" and "pleasure". This pattern helps us realize that fan fiction writers care much more about romantic/sexual interactions than any specific character.

Fan fiction stories in Chinese and English share all patterns mentioned above but also differ in some ways. Aside from a bit different tastes on types of character pairings, Chinese and English stories also use slightly different languages to write about sex scenes. In Chinese stories, there are more sensation words such as "xuàn yùn(whirl)" and "chén nì(overwhelm)", as well as reaction words such as "chàn dǒu(shiver)" and "mǎn zú(slake)". In English, there are mostly direct descriptions with body movement words like "fuck", "groan" and "touch." We also notice that the English word "fuck" is frequently used in Chinese stories. Instead of "cāo" which is exactly what the word means in Chinese, an English counterpart might be less shameful to mention and thus makes writers and readers more comfortable. This difference in language use implies people's openness to talk and write about sex in different language cultures. Writing and reading fan fiction stories as a secret way to fulfill sexual desire, people using Chinese tend to feel more shameful even though in the safe closed fan fiction community, and thus prefer indirect languages. This might also connect to our findings on lexical diversity. By calculating type-token ratio (TTR) for stories on each canon by languages, we notice that Chinese stories are generally richer in the variety of vocabulary. One potential explanation for it could be the habit of indirect expressions in Chinese stories that bring about creativity and uniqueness in terms of word selections.

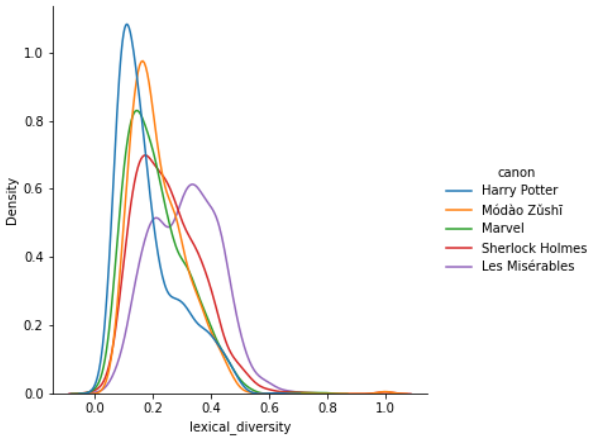

Figure 13: Lexical Diversity of stories in English: KDE density plot

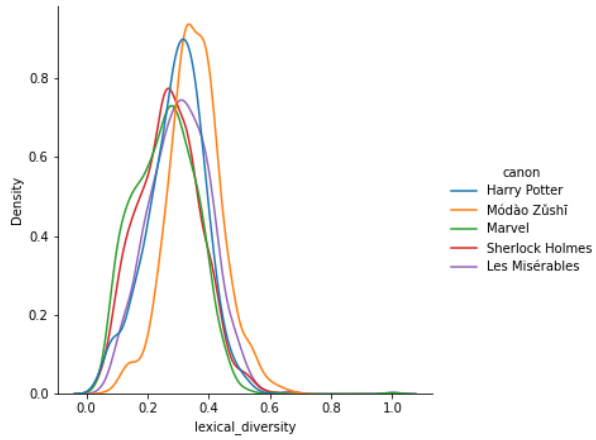

Figure 14: Lexical Diversity of stories in Chinese: KDE density plot

Discussion

Based on all of our findings, the primary focus on male homosexuality in fan fiction stories is actually a female version of the "male gaze", which puts a male character's form (sexiness and "to-be-looked-at-ness") over his content (character development). Writing male-male romance empowers females to distance "sex" from their own bodies and gain to power to watch, desire, and abuse males freely. It is silent revenge of women against the real patriarchal world as the unilaterally consumed and sexually suppressed.

At the same time, perspectives in fan fiction stories are also socialized. Judith Butler in *Critically Queer* ((Butler, 1993)) argues that gender performativity is "compulsory repetitions of prior and subjectivizing norms...which work, animate, and constrain the gendered subject." Writers usually rely on their female heterosexual experiences to write about a man in a relationship with another man. Thus, male homosexuality in fan fiction stories usually mimics modes of male-female romance, and sex scenes usually emphasize feelings of the bottom more. Socialized perspectives also differ by language cultures. The guilt and shame culture of the Chinese language makes Chinese writers select less explicit vocabulary and more indirect expressions for sex scenes.

Acknowledgments

We thank Professor David Bamman and Professor Tom McEnaney for their helpful comments and feedback, which helped us to greatly improve the paper.

References

Camille Bacon-Smith. 1991. *Enterprising Women: Television Fandom and the Creation of Popular Myth*. University of Pennsylvania Press, Philadelphia, PA.

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks.

Judith Butler. 1993. Critically queer. *GLQ*, 1:17–32.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Henry Jenkins. 1992. *Textual Poachers: Television Fans and Participatory Culture*. Routledge, New York, NY.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.*, 52(5).

Smitha Milli and David Bamman. 2016. Beyond canonical texts: A computational analysis of fanfiction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Laura Mulvey. 1975. Visual pleasure and narrative cinema. *Screen*, 16:6–18.

Muon Nguyen, Thanh Ho, and Phuc Do. 2013. Social networks analysis based on topic modeling. In *The 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 119–122.

Constance Penley. 2014. *The Fan Fiction Studies Reader*. University of Iowa Press, Iowa City.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Gaye Tuchman and Nina Fortin. 1989. *Edging Women Out*. Routledge, New York, NY.