Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own

Thomas Schmidt, Johanna Grünler, Nicole Schönwerth & Christian Wolff

Media Informatics Group, University of Regensburg, Germany

{firstname.lastname@ur.de}

2nd International Conference of the European Association for Digital Humanities (EADH 2021)
Krasnoyarsk, Russia
September 21-25, 2021

Keywords: fan fiction, online writing, digital humanities, literary studies, fan studies, online communities, corpus creation, corpus analysis

Abstract.
We report upon a digital humanities project on the acquisition and analysis of a corpus of German online writings. We have implemented a scraper to gather the German language material as well as corresponding metadata of the popular online writing platform Archive of Our Own (AO3), which is a platform primarily focused on the text sort of fan fictions. The corpus consists of 9,640 writings resulting in over 39 million tokens and 3.6 million sentences. The texts have varying lengths with a median of around 2,500 tokens per story. We present results on the analysis of metadata and general text statistics like the most frequent words. While we can support previous findings of literary and media studies like the dominance of male-male romantic and erotic narratives, we can also identify attributes that are very specific and unique to German culture as well as differences to results of research for English online writings. We will outline in our future work how we plan to further increase and analyze the corpus to support research in digital humanities as well as German literary and fan studies.

Cite as:

Schmidt, T., Grünler, J., Schönwerth, N. & Wolff, C. (2021). Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own. In *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*. Krasnoyarsk, Russia.

1. Introduction

Online media and content have gained a lot of interest in *Digital Humanities* (DH) in recent years (e.g. Moßburger et al. 2020; Schmidt et al. 2020a; Schmidt et al. 2020c). In the context of literary studies, the analysis of online creative writing platforms has gained more and more

popularity (Hellekson / Busse 2006; Jamison 2013). While some platforms focus on the creation of original content, other platforms like *Archive of our Own* (AO3)[1] and *Fanfiction.net*[2] focus on the specific genre of fan fiction. Fan fictions are fan-created works using already existing characters and plot elements of existing famous media like literature, movies or games to write new stories based on those characters (Dym et al. 2018). Scholars have analyzed the history and cultural influence of this text genre (Cuntz-Leng / Meintzinger 2015; Hellekson / Busse 2006; Jamison 2013; Thomas 2011; Van Steenhuyse 2011). Hellekson and Busse (2006) highlight the striking dominance of slash fan fiction (stories focused on male-male romantic and erotic relationships) in the fan fiction community. Researchers in *Natural Language Processing* (NLP) make use of the online availability of these large bodies of narrative texts with rich metadata to explore and evaluate new methods (Liu et al. 2019; Muttenthaler et al. 2019; Vilares / Gómez-Rodríguez 2019; Zhang et al. 2019). However, fan fictions themselves have also been subject to computational research. Among other, researchers examine the metadata of fan fictions (Milli / Bamman 2016; Yin et al. 2017; Kleindienst / Schmidt 2020), gender and stereotypes (Fast et al. 2016) and the role and content of user feedback (Frens et al. 2018; Pianzola et al. 2020; Rebora / Pianzola 2018).

In general, the focus of research is currently on English fan fictions. However, researchers in humanities argue that style, content and progression of fan fictions differ with respect to different regional cultures (Cutz-Leng / Meintzinger 2015). We propose that country-specific features, which are of interest for DH as well as cultural studies, might be expressed in such corpora and should be analyzed to verify such assumptions. Therefore, we want to investigate the benefits of the corpus analysis of non-English fan fiction for the example of German. For the preliminary analysis presented in this abstract, we focus on metadata of fan fiction and how it reflects national-specific features. In future studies we plan to compare the content of multiple languages to each other to identify country specific differences in content and style.

2. Corpus

We have chosen AO3 as the source for creating our preliminary corpus. AO3 describes itself as a non-commercial archive for transformative fan fiction.

We have created a scraper to gather every chapter of every German text on AO3 (more precisely texts marked as German by the creator) and the corresponding metadata by using the language-based search function of AO3. AO3 explicitly allows the scraping of their content in their terms of use. Filtering AO3 for languages shows that 93% of all texts are marked as English, while only 7% are non-English. Overall, the German texts account for 0.2% of all AO3 material only. We acquired the German texts in September 2019. We filtered out any non-German text as well as pages containing solely links, pictures or text pages that were empty. This reduced the overall number of writings to 9,640[3].

[1] https://archiveofourown.org/
[2] https://www.fanfiction.net/
[3] Due to legal issues the corpus is currently only available upon request via mail (thomas.schmidt@ur.de). We will publish parts of the corpus via the following GitHub repository: https://github.com/lauchblatt/German_Fan_Fictions

Next to the text, AO3 offers a rich set of metadata which is currently in the focus of our analysis. Table 1 summarizes the attributes of the items of the corpus. Table 2 illustrates the basic statistics of the corpus. Tokenization was performed via the *NLTK* standard tokenizer[4] and sentence splitting via NLTK and the *Punkt* sentence splitter[5].

Key	**Value description**
author	username of the author
title	the title of the work
text	the entire text of all the chapters of a work
category	a tag to illustrate the romantic or sexual relation displayed in the story e.g., M/M for a male-on-male relationship
rating	indicates if the story contains any sensitive material by addressing the audience type
archive_ warning	a warning authors can use to inform that the story might contain sensitive material e.g., “Major Character Death” or “Graphic Depictions of Violence”
fandom	the fandom a story is about (e.g., “Harry Potter”) or the information that this is an original work
character	list of characters that appear in the story
relationship	description of which character has a relationship with another one
additional_tag	list of additional tags an author might add

Table 1. Structure of a corpus item.

Metric	**Tokens**	**Sentences**
Total	79,316,704	3,662,344
Min	3	1
Avg	11,060.2	513.92
Med	2,672	119
Max	2,312,247	105,362
Std	39,001.6	1,787.3

Table 2. Token and sentence statistics.

3. Metadata analysis

We present and discuss results about metadata analysis that show specific expressions of German culture and therefore allow us to further investigate differences and features in online writings and fan culture. To identify nation-specific differences we compared results of our corpus to research on English-dominated corpora (Milli / Bamman 2016; Yin et al. 2017) as well as on fan-based analysis on AO3 in general[6].

[4] https://www.nltk.org/api/nltk.tokenize.html
[5] http://www.nltk.org/_modules/nltk/tokenize/punkt.html
[6] For more information visit: https://destinationtoast.tumblr.com/post/157728590234/toastystats-top-fandoms-on-ao3-as-of-february-26

We found that many of the most popular fandoms are indeed specific expressions of German culture (table 3). The most popular one being *Tatort*: a German Sunday evening police procedural television series. Fan fictions based on real persons from popular sports in Germany like soccer and ski jumping are also rather popular in Germany. Taking the most popular relationships into account, we also found that stories about the two famous German poets *Schiller* and *Goeth*e are quite frequent (table 6). Other than that, fandom distributions are similar to other research (Milli / Bamman 2016; Yin et al. 2017) with *Harry Potter*, *Supernatural* and *Sherlock* being among the most popular fandoms. It is often argued that the rise of fan fictions is strongly intertwined with Anime in Germany (Cuntz-Leng / Meintzinger 2015); however, in our corpus the most popular Anime-fandom is *Naruto* with only 97 stories showing that Anime is not as popular as in general on AO3.

Fandom	Frequency	Percentage
Tatort	986	10.2%
Harry Potter	800	8.3%
Supernatural	413	4.3%
Sherlock (TV)	405	4.2%
Original Work	349	3.6%
Football RPF	295	3.1%
Stargate Atlantis	220	2.3%
Stargate SG	191	2.0%
Historical RPF	151	1.6%
Glee	141	1.5%
Teen Wolf (TV)	138	1.5%
The Avengers	133	1.4%
Ski Jumping RPF	131	1.4%
Rest (1603 Fandoms)	5,287	54.8%

Table 3. Distribution of fandoms.

One of the most striking attributes of the corpus is the dominance of male-male relationships (table 4) and male characters in general as shown by the analysis of most popular characters (table 5), which are predominantly male, and the most popular relationships which are all male (table 6). The popularity of this type of stories is a well-documented attribute of fan fictions (cf. Hellekson / Busse 2006). Please note that this content does not have to be erotic or sexualized but is mostly focused on romance and friendship as can be seen in the analysis of additional tags (table 7). While research in the humanities focuses on explaining this popularity via gender and political discourse (Duggan 2017; Hellekson / Busse 2006; Tosenberger 2008), we also plan to support this research with computational methods.

Relationship	Frequency	Percentage
F/F (Female-female)	386	4%
F/M (Female-male)	1,906	20%
M/M (Male-male)	5,429	56%
Multi	262	3%
Gen (General, mostly meaning that relationships are not too important)	2,052	21%
Other	201	2%

Table 4. Distribution of relationship categories.

Character	Fandom	Frequency
Karl-Friedrich Boerne	Tatort	845
Frank Thiel	Tatort	801
Sherlock Holmes	Sherlock (TV)	366
John Watson	Sherlock (TV)	348
Harry Potter	Harry Potter	341
Dean Winchester	Supernatural	318
Severus Snape	Harry Potter	268
Draco Malfoy	Harry Potter	255
Original Characters	Original Work	254
Sam Winchester	Supernatural	253
Hermione Granger	Harry Potter	221
Original Female Character(s)	Original Work	190
John Sheppard	Stargate Atlantis	186
Original Male Character(s)	Original Work	185

Table 5. Distribution of the most frequent character tags.

Relationships	Fandom	Frequency
Karl-Friedrich Boerne / Frank Thiel	Tatort	827
Sherlock Holmes / John Watson	Sherlock (TV)	367
Castiel / Dean Winchester	Supernatural	165
Harry Potter / Draco Malfoy	Harry Potter	145
Blaine Anderson / Kurt Hummel	Glee	122
Johann Wolfgang von Goethe / Friedrich Schiller	Historical Person Fiction	115
Daniel Jackson / Jack O'Neill	Stargate SG-1	103
Rodney McKay / John Sheppard	Stargate Atlantis	95
Derek Hale / Stiles Stilinski	Teen Wolf (TV)	90
Mycroft Holmes / Greg Lestrade	Sherlock	84

Table 6. Distribution of the 10 most frequent character relationships.

Additional Tags	Frequency	Percent
Deutsch | German	1,472	10.6%
Fluff	940	6.8%
Humor	685	4.9%
Romance	580	4.2%
Friendship	572	4.1%
Hurt/Comfort	511	3.7%
Angst	483	3.5%
Male Slash	365	2.92%
Established Relationship	352	2.5%
Drama	341	2.5%

Table 7. Distribution of the 10 most frequent additional tags.

While we focus on metadata analysis in this paper, we also have performed some basic text analyses. Table 8 illustrates the most frequent words of the entire corpus after stop word removal. Striking is the rather frequent usage of terms describing physical attributes (*augen*, *hand*, *kopf*, *gesicht*, *stimme*). Since the data considering the metadata show that most stories are relationship- and romance-driven, we assume that those terms point to romantic and erotic descriptions and actions of the characters.

References

Buhl, H. (2013). *Tatort: gesellschaftspolitische Themen in der Krimireihe*. UVK.

Cuntz-Leng, V., & Meintzinger, J. (2015). A brief history of fan fiction in Germany. *Transformative Works and Cultures*, 19.

Duggan, J. (2017). Revising hegemonic masculinity: Homosexuality, masculinity, and youth-authored Harry Potter fan fiction. *Bookbird: A Journal of International Children's Literature*, 55(2), 38-45.

Dym, B., Aragon, C., Bullard, J., Davis, R., & Fiesler, C. (2018, October). Online Fandom: Boldly Going Where Few CSCW Researchers Have Gone Before. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 121-124). ACM.

Fast, E., Vachovsky, T., & Bernstein, M. S. (2016b, March). Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Tenth International AAAI Conference on Web and Social Media.*

Frens, J., Davis, R., Lee, J., Zhang, D., & Aragon, C. (2018). Reviews Matter: How Distributed Mentoring Predicts Lexical Diversity on Fan fiction. *arXiv preprint arXiv*:1809.10268.

Hellekson, K. & Busse, K. (2006). *Fan Fiction and Fan Communities in the Age of the Internet: New Essays*. Jefferson, NC: McFarland.

Jamison, A. (2013). *Fic: Why fan fiction is taking over the world*. BenBella Books, Inc.

Liu, C., Osama, M., & De Andrade, A. (2019). DENS: A Dataset for Multi-class Emotion Analysis. *arXiv preprint arXiv*:1910.11769.

Milli, S., & Bamman, D. (2016, November). Beyond canonical texts: A computational analysis of fan fiction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2048-2053).

Moßburger, L., Wende, F., Brinkmann, K., & Schmidt, T. (2020, December). Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task* (pp. 70-81).

Muttenthaler, L., Lucas, G. & Amann, J. (2019). Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum.*

Painter, D. T., Daniels, B. C., & Jost, J. (2019). Network analysis for the digital humanities: principles, problems, extensions. *Isis*, 110(3), 538-554.

Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PloS one*, 15(1), e0226708.

Rebora, S., & Pianzola, F. (2018). A New Research Programme for Reading Research: Analysing Comments in the Margins on Wattpad. *DigitCult-Scientific Journal on Digital Cultures*, 3(2), 19-36.

Schmidt, T. & Burghardt, M. (2018a). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 139-149). Santa Fe, New Mexico: Association for Computational Linguistics.

Schmidt, T. & Burghardt, M. (2018b). Toward a Tool for Sentiment Analysis for German Historic Plays. In: Piotrowski, M. (ed.), *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018* (pp. 46-48). Lausanne, Switzerland: Laboratoire laussannois d'informatique et statistique textuelle.

Kleindienst, N. & Schmidt, T. (2020). Investigating the Transformation of Original Work by the Online Fan Fiction Community: A Case Study for Supernatural. In *Digital Practices. Reading, Writing and Evaluation on the Web.* Basel, Switzerland.

Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. (2019a). Katharsis - A Tool for Computational Drametrics. In: *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands. http://dx.doi.org/10.5283/epub.43579

Schmidt, T., Burghardt, M. & Wolff, C. (2019b). Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)* (pp. 405-414). Copenhagen, Denmark.

Schmidt, T., Hartl, P., Ramsauer, D., Fischer, T., Hilzenthaler, A. & Wolff, C. (2020a). Acquisition and Analysis of a Meme Corpus to Investigate Web Culture. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts.* Ottawa, Canada. http://dx.doi.org/10.17613/mw0s-0805

Schmidt, T., Bauer, M., Habler, F., Heuberger, H., Pilsl, F. & Wolff, C. (2020b). Der Einsatz von Distant Reading auf einem Korpus deutschsprachiger Songtexte. In *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* (pp. 296-300). Paderborn, Germany. http://dx.doi.org/10.5281/zenodo.4621928

Schmidt, T., Kaindl, F. & Wolff, C. (2020c). Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)* (pp. 157-172). Riga, Latvia.

Schmidt, T., Kaindl, F. & Wolff, C. (2020d). Visualizing Collocations in Religious Online Forums. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts.* Ottawa, Canada. http://dx.doi.org/10.17613/aq1q-1t69

Schöch, C. (2021). Topic modeling genre: an exploration of french classical and enlightenment drama. *arXiv preprint arXiv*:2103.13019.

Sprugnoli, R., Tonelli, S., Marchetti, A., & Moretti, G. (2016). Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4), 762-772.

Thomas, B. (2011). What Is Fan fiction and Why Are People Saying Such Nice Things about It??. *Storyworlds: A Journal of Narrative Studies*, 3, 1-24.

Tosenberger, C. (2008). Homosexuality at the online Hogwarts: Harry Potter slash fan fiction. *Children's Literature*, 36(1), 185-207.

Van Steenhuyse, V. (2011). The writing and reading of fan fiction and transformation theory. *CLCWeb: Comparative Literature and Culture*, 13(4), 4.

Vilares, D., & Gómez-Rodríguez, C. (2019). Harry Potter and the Action Prediction Challenge from Natural Language. *arXiv preprint arXiv*:1905.11037.

Yin, K., Aragon, C., Evans, S., & Davis, K. (2017, May). Where No One Has Gone Before: A Meta-Dataset of the World's Largest Fan fiction Repository. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 6106-6110). ACM.

Zhang, W., Cheung, J. C. K., & Oren, J. (2019, July). Generating Character Descriptions for Automatic Summarization of Fiction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 7476-7483).