From Mining to Profiling: A computational literary case study on gender features in literary texts

Short description:
In our case study, we explore overarching patterns of characterization in a corpus of narratives. From each of the 19 texts in our corpus, we closely read and annotated features used for the characterization of literary characters in the first 20.000 tokens. Thus, we focused on the introduction of characters and the establishment of social character networks. Using this information we build graphs with character profiles. We found that every character is connected to multiple features and that these features are most often distinctively used for and more rarely shared by a variety of characters. No typical gendered features show. Nevertheless, we were able to build four groups of social setups based on characterization.

Long Description:
In Computational Literary Studies (CLS) research that considers gender aspects in relation to the representation of literary characters is slowly becoming more frequent. In some studies, gender is one trait or feature of the literature analyzed as one case study in the course of others, e.g. in Piper (2019) or Underwood (2019). Gender has been analyzed in terms of agency (Jockers and Kiriloff 2016) or action (Baylog et al. 2014, Kiriloff and Jockers 2018) as well as in relation to emotions (Flüh and Schumacher 2022). The (under)representation of female protagonists has been focused on by Kraicer and Piper (2019). Gender has not only been of interest in terms of characterization but also in regard to authorship and style (e.g. Rybicki 2015, Weitin 2021). However, it has been rightfully criticized that many Digital Humanities studies used a basic understanding of gender as a binary phenomenon (Mandell 2019). While in the natural, physical and social sciences, the inclusion of feminist theories and methods has already been done, this inclusion is rare in CLS. Nevertheless, the advantages are already becoming clear: The value of an inclusion into CLS can be seen in the fact that deep-seated and anchored conceptions about gender can be broken down through feminist-oriented data mining (Rhody 2016). This finding fits the recently observed bridging between Digital Humanities (DH) and cultural studies, in which in addition to 'gender' and 'race', critical and feminist theory also strives to intertwine with the DH (Deremetz 2023, 34 pp).

Despite being an integral part of most literary texts, features attributed to characters have not yet been systematically analyzed in terms of gender. In their role as cultural objects, literary texts reproduce and shape, the perspective on characteristics judged to be appropriate or inappropriate. In our case study, we explore the question of which overarching patterns of characterization show in a corpus of narratives. In terms of gender, one can find an extensive analysis of binary and non-binary gender roles in Beauvoir's text *the second sex* (1949). Her assumptions are based on psychological and social science research as well as on the study of literature. She mentions more than 100 primary sources of relevance for gender-focused literary studies We extracted the literary texts mentioned by de Beauvoir by close reading (the entire list is available here). To distinguish tendencies of characterization and (non)binarity of gender, we present the results of a corpus-based analysis of a subcorpus of these texts.

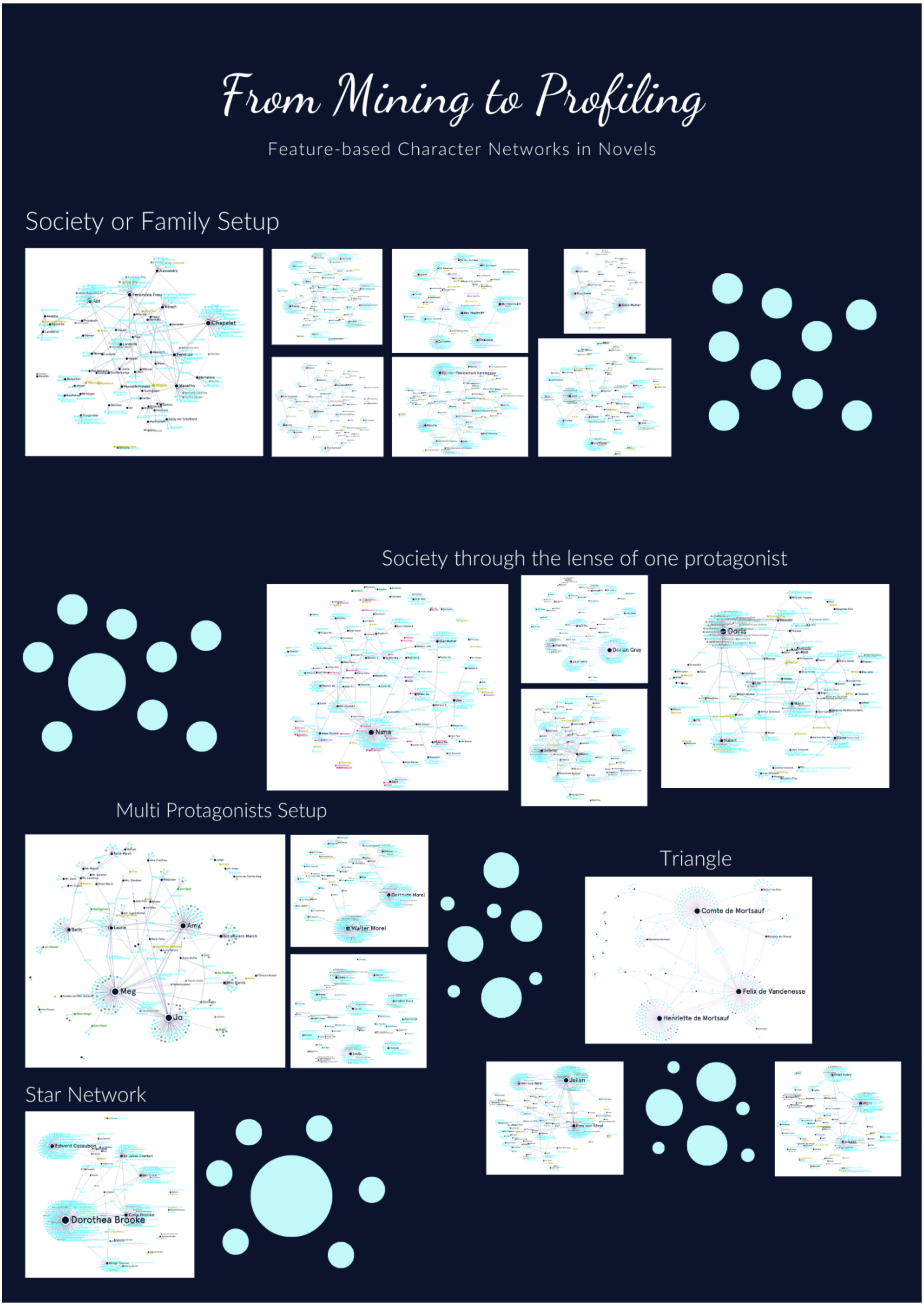
Method

Our corpus contains 19 novels altogether. In turn of a pragmatic approach, we closely read and annotated the first 20.000 tokens from every text in order to focus on the introduction and establishment of characters. We annotated features, which we understand as long-lasting characteristics of literary characters. Furthermore, we annotated descriptions of clothing. Using this information we build graphs with character profiles. To each character, we additionally assigned one most dominant gender role. Then we interpreted the data by text (meaning the first 20.000 tokens) and corpus (meaning 380.000 tokens from 19 different texts). Based on the distanced view of all networks, we derived different main types of network graphs with recurrent patterns.

Selected Findings

Based on the features used to describe and establish characters we split up our corpus into five groups with prototypical patterns (cf. figure 1)

1. **Society or Family Setup**: many characters are described with many features. Usually, protagonists include a more complex description than minor characters but they are closely followed by other major characters.
2. **Society through the lens of the protagonist**: This group shows similar networks in size and degree of connectedness as type 1, but protagonists are more dominant here and clearly show the most (mainly distinctive) features.
3. **Multi-protagonist setup**: more than one character is described by a whole range of (mainly distinctive) features.
4. **Triangle**: a special case of a multi-protagonist setup in which three protagonists show.
5. **Star Network**: the protagonist is described with a great number of features.



From Mining to Profiling

Feature-based Character Networks in Novels

Society or Family Setup

Society through the lense of one protagonist

Multi Protagonists Setup

Triangle

Star Network

figure 1: Graphs grouped by patterns of characterization per text with schematic illustrations of different network types

When taking the whole corpus into view, some more characteristics of the analyzed narratives in terms of character descriptions show (cf. figure 2). A whole range of distinctive features is used for characterization whereas features shared by many characters are rather rare and they are often undefined in terms of gender. The five most common shared features are ‘young‘, ‘small‘, ‘old‘, ‘poor‘ and ‘beautiful‘ (‘beautiful‘ does show a bias towards feminine characters but is nevertheless frequently used for male and non-binary gender roles.). It also shows that the binary gender roles of men and women are indeed the ones that the characters often primarily stand for. In addition, there are more characters altogether that are related to a gender role of the male spectrum (327 characters) than to the female spectrum (241 female characters), which matches the findings by Kraicer and Piper (2019).

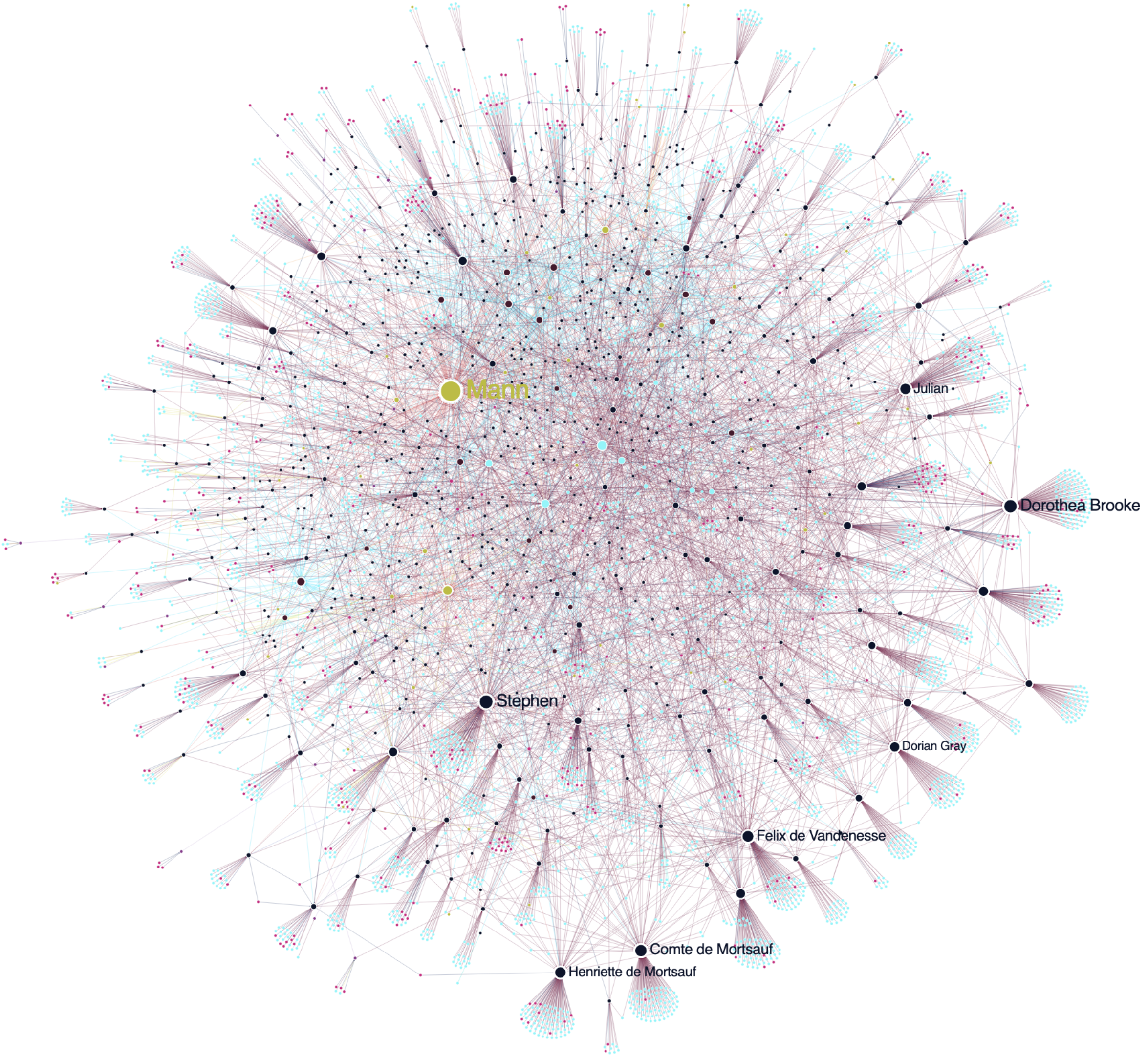


figure 2: Graph showing characterizations in the entire corpus

Conclusion

Acknowledging the fact that there still is some basic research to be done on typical descriptions of gender in literature in order to find out when individuality and diversity start and that binarity should be left behind in favor of a more diverse understanding of gender, we developed a mode of analysis to which complex, graph-based character profiles are central. Using this method mainly two things show: (1) every character is connected to multiple features and (2) these features are most often distinctively used for and more rarely shared by a variety of characters. More findings, especially regarding distinctive features of a variety of gender roles will be presented in our talk.

Bibliography

Baylog, O. *et al.* (2016) ‘" More than Custom has Pronounced Necessary” : Exploring the Correlation between Gendered Verbs and Character in the 19 th Century Novel Nebraska Literary Lab’, in. Available at: https://www.semanticscholar.org/paper/%22-More-than-Custom-has-Pronounced-Necessary-%E2%80%9D-%3A-the-Baylog-Dimmit/631b70f5581d6df8b1e35f9f679465a119b499ba (Accessed: 12 January 2023).

Beauvoir, S. (1992). *Das andere Geschlecht Sitte und Sexus der Frau*. Neuübers. Rowohlt-Taschenbuch-Verl. Hamburg: Rowohlt.

Deremetz, A. (2023). *Mixed Methods in den Digital Humanities. Topic-informierte Diskursanalyse am Beispiel der Volkszählungs- und Zensusdebatte*. Stuttgart: Metzler.

Flüh, M., Schumacher, M. K. (2022). Jung, wild, emotional? Rollen und Emotionen Jugendlicher in zeitgenössischer Fantasy-Literatur”. In: *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2022),* Potsdam. https://doi.org/10.5281/zenodo.5555952.

Jockers, M. and Kirilloff, G. (2016) ‘Understanding Gender and Character Agency in the 19th Century Novel’, *Journal of Cultural Analytics*, 2(2). Available at: https://doi.org/10.22148/16.010.

Kirilloff, G. *et al.* (2018) ‘From a distance “You might mistake her for a man”: A closer reading of gender and character action in Jane Eyre, The Law and the Lady, and A Brilliant Woman1’, *Digital Scholarship in the Humanities*, 33. Available at: https://doi.org/10.1093/llc/fqy011.

Kraicer, E. and Piper, A. (2019) ‘Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction’, *Journal of Cultural Analytics*, 3(2). Available at: https://doi.org/10.22148/16.032.

Mandell, L. (2019) ‘Gender and Cultural Analytics:: Finding or Making Stereotypes?’, in, pp. 3–26. Available at: https://doi.org/10.5749/j.ctvg251hk.4.

Piper, A. (2018) *Enumerations: data and literary study*. Chicago ; London: The University of Chicago Press.

Rhody, L. M. (2016). Why I dig: Feminist Approaches in the Digital Humanities. In *Debates in the Digital Humanities*. https://dhdebates.gc.cuny.edu/read/untitled/section/508c8664-15c8-4262-a72a-e49299873d11 Cite [Access: 20th February 2023].

Rybicki, Jan (2015). “Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies.” In *Digital Scholarship in the Humanities*: 1–16. doi: 10.1093/llc/fqv023.

Underwood, T., Bamman, D. and Lee, S. (2018) ‘The Transformation of Gender in English-Language Fiction’, *Journal of Cultural Analytics*, 3(2). Available at: https://doi.org/10.22148/16.019.

Weitin, T. (2021) *Digitale Literaturgeschichte: Eine Versuchsreihe mit sieben Experimenten*. Berlin, Heidelberg: Springer (Digitale Literaturwissenschaft). Available at: https://doi.org/10.1007/978-3-662-63663-3.

Corpus

Alcott, Louisa May (1868/69). *Kleine Frauen* (Little Women)
Balzac, Honoré de (1835). *Die Lilie im Tal* (The Lily of the Valley)
Boccaccio, Giovanni (1353). *Das Decameron* (1–6) (The Decameron)
Brontë, Emily (1847). *Sturmhöhe* (Wuthering Heights)
Chateaubriand, François-René de (1801). *Atala*
Colette, Sidonie-Gabrielle Claudine (1929). *Sido*
Dostojewski, Fjodor Michailowitsch (1880). *Die Brüder Karamasow* (The Karamazow Brothers)
Eliot, George (1871). *Middlemarch*
Keun, Irmgard (1932). *Das kunstseidene Mädchen* (The Artificial Silk Girl)
Hall, Radclyffe (1928). *Quell der Einsamkeit* (The Well of Loneliness)
Lawrence, David Herbert (1913). *Söhne und Liebhaber* (Sons and Lovers)
Sade, Marquis de (1797). *Juliette*
Steinbeck, John (1945). *Die Straße der Ölsardinen* (Cannery Row)
Stendhal, Henry-Marie Beyle (1830). *Rot und Schwarz* (The Red and the Black)
Tolstoi, Lew Nikolajewitsch (1867). *Krieg und Frieden* (War and Peace)

Wilde, Oocar(1890). *Das Bildnis des Dorian Gray* (The Picture of Dorian Gray)
Woolf, Virginia (1931). *Die Wellen* (The Waves)
Zola, Émil (1880). *Nana*
Zola, Émil (1882). *Ein feines Haus* (Pot Luck)