The FairyNet Corpus - Character Networks for German Fairy Tales

David Schmidt, Albin Zehe, Janne Lorenzen, Lisa Sergel, Sebastian Düker, Markus Krug, Frank Puppe
Universität Würzburg, Germany
firstname.lastname@informatik.uni-wuerzburg.de

Abstract

This paper presents a data set of German fairy tales, manually annotated with character networks which were obtained with high inter-rater agreement. The release of this corpus provides an opportunity of training and comparing different algorithms for the extraction of character networks, which so far was barely possible due to heterogeneous interests of previous researchers. We demonstrate the usefulness of our data set by providing baseline experiments for the automatic extraction of character networks, applying a rule-based pipeline as well as a neural approach, and find the neural approach outperforming the rule-approach in most evaluation settings.

1 Introduction

The creation and publication of data sets that can be shared without restrictions is a driving factor in the progress of digital humanities. In this paper we describe a new resource, comprising 40 German fairy tales by the Brothers Grimm. Our corpus contains the annotation of all fictional entities and their references as well as a manually created character network for each of the texts. The main contribution of this work is a publicly available resource[1] that serves the purpose of supporting the creation of automatic algorithms for character network extraction. The paper is structured as follows: First, we give a small overview of the most prominent works that deal with automatic extraction of character networks as well as the very heterogeneous ways the authors evaluate the networks. We then describe the data and its origin, our guidelines and the inter-rater agreement we achieved on the different layers of annotation in more detail in section 3. In section 4 we present preliminary experiments that can be referred to as baselines in future projects using the corpus, before concluding the paper.

[1] https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/fairynet-latech-clfl

2 Related Work

Since to the best of our knowledge no freely available data set comprising character networks for prose texts that originate a standardized set of guidelines exist, we cover the different approaches that dealt with the automatic extraction of character networks and depict how they went around this issue of a missing data set. For a wide overview of this field we refer to Labatut and Bost (2019). For textual media, the methods to create character networks share many similarities, especially in regard to the lexical preprocessing involved. The similiarities end, when we compare how character networks are evaluated. The field is still very heterogeneous, rendering direct comparisons of different approaches almost infeasible. By publication of this data set we hope to address this point and help to support convergence in terms of character network evaluation.

Most previous work only evaluates their extracted networks by proxies. The system of Park et al. (2013) evaluate by assuming the character and co-occurence frequencies (between characters) are distributed by a power law, which they were able to recover. The work of Dekker et al. (2019) uses modern and classical literature and compare their character networks using Social Network Analysis (SNA), but found that there is no significant difference between these two sets. The work of Coll Ardanuy and Sporleder (2014) extracted networks and subsequently clustered them to examine whether the networks can be used as proxies for unsupervised genre and authorship attribution, but could only slightly exceed their baselines. The work of Elson et al. (2010) can be seen as one of the first papers that dealt with the automatic extraction of character networks. They validated their networks by the verification of different literary theories (e.g. they suspected that the more characters are involved in a story, the less dialogue is con-

Proceedings of LaTeCH-CLfL 2021, pages 49–56
Punta Cana, Dominican Republic (Online), November 11, 2021.

tained). Agarwal et al. (2012) and Jayannavar et al. (2015) extracted interactions between characters using relations of the categories *interact* and *observe* which they learn using a kernel Support Vector Machine. They evaluated their system based on SNA metrics, such as the Node-In-Degree-Centrality or Out-Centrality. The method of Jing et al. (2007) evaluated the networks by measuring the overlap between nodes and edges. Their results suggest that coreference resolution is the limiting factor of their approach, only reaching F1-scores of 0.3. The work of Krug (2020), extracted gold character networks from labeled expert summaries of novels, and compared the gold networks with system networks extracted from the full text novels using ranking metrics.

3 Data and Annotation

The texts that we plan to annotate are the fairy tales of the seventh edition of the *Children's and Household Tales* by the Brothers Grimm for the German language[2]. Currently, we finished annotating 40 fairy tales alongside their character networks. Each fairy tale now consists of two files: The first file is the text (modernized with a tool from Deutsches Textarchiv[3] (Jurish, 2012)) with markup representing the references of the characters (including coreference annotations) and the second file contains a manually annotated character network in the form of a ".xlsx" file (for automatic usability we will release .json as well), containing several layers of annotations, described in the following section.

3.1 Annotation Guidelines

This section describes the guidelines that were used to annotate the different layers of the annotations.

3.1.1 Texts

The characters and coreferences were annotated simply by their heads in a fashion similar to that of DROC (Krug et al., 2018), instead of marking complete noun phrases like in Ontonotes (Weischedel et al., 2011)[4]. On top of the three **syntactic categories** of a mention (name, noun phrase, pronoun), we also attributed each mention with a **semantic category** ("human", "legendary creature", "transcendent creature" and "generic"). The semantic category "human" is attached to all human character references, while the category "legendary creature" is attached to all references of entities which show some inhuman traits (e.g. a speaking animal, a sorcerer, dead people showing unexpected behaviour, giants or dwarves.). A transcendent character is reserved for religious entities that do not actually appear in the story but are referenced in sayings (e.g. "god" or "devil"). The last category is a "generic" reference. It is used for mentions which do not refer to a concrete entity (e.g. "[nobody] could have saved him"). This label overrides potential labels "human" or "legendary creature" if the character is considered to be abstract.

We also annotate **grammatical number** and **biological sex** (i.e., a reference to a girl with the German neuter pronoun "es" (it), as in "das Mädchen" would be marked as a singular female).

3.1.2 Networks

Each character that plays an important role in the plot is deemed relevant enough to be in the final network of the story (linked via their coreference ID into the text file) and is assigned one of the **character types** of Vladimir Propp (1972) (e.g. hero, villain, etc..).[5] Examples for characters that are not relevant enough for the networks are the cook and the kitchen boy in *Dornröschen* (*Sleeping Beauty*) because their existence and actions have no influence on the plot whatsoever. In cases where only the actions of a group of characters but not those of the individual characters are important for the plot, the networks contain only a node for the group (e.g. the dwarves in *Schneewittchen* (*Snow White*) or the parents of the princess in *Dornröschen* (*Sleeping Beauty*).

For the **age** of the entities, we use a set of five distinct labels: (1) Baby: should be assigned to newborns only (2) Growing-up: An entity that is either a child or adolescent. (3) Grown-up: Entities that take on responsibility on their own (having children is one such responsibility). (4) Aged: Requires a clear mention in the text ("alte Hexe"/"old witch", "Großvater"/"grandfather" - here we assume third generation people to take the value by default) (5) Unknown: This category is used if there are no clear indicators, for instance for legendary creatures.

The **entity sentiment** of a character corresponds

[2]https://de.wikisource.org/wiki/Kinder-_und_Hausmärchen

[3]https://www.deutschestextarchiv.de/demo/cab/

[4]Example: In the noun phrase "eine kleine süße Dirne" (a little sweet girl) we only mark "Dirne" (girl).

[5]We restricted the networks to the important characters to prevent them from being cluttered and to keep the amount of work manageable.

to the sentiment of the entity as it is depicted by the *author*. We only make use of three possible values for the sentiment of a character: (1) positive: This value is assigned if the entity is described to be social and acts in an ethical manner (e.g. the entity cares for the well-being of someone else) and/or is attributed positive character traits, e.g. "Das Mädchen [...] blieb fromm und gut." (The girl [...] stayed pious and good.) (2) negative: This value is assigned if the character only cares about themselves and/or acts unsocial and unethical and/or is attributed negative character traits, e.g. "eine böse Hexe". (3) neutral: By default, a character receives this sentiment value. The sentiment can change over the course of the plot (though it only happened once in the 40 documents we have annotated).

The edges in the network correspond to the sentiment between characters and the social relations: By analyzing all thoughts an entity has towards another entity as well as all actions taken between them, we are trying to subsume the **sentiment between characters** in a label with the values: (1) positive: Entity has either positive feelings, thoughts or had actions that she would repeat. (2) negative: Entity has negative feelings (such as fear), negative thoughts or fear of reliving certain actions with the other entity (e.g. being eaten by the other entity). (3) neutral (4) none/empty cell: There is no mention of thoughts or interaction between the characters. The set of **social relations** is currently spanning 29 labels, with more to be expected to come if we continue to label more texts. These contain family relations (16 labels), a label that expresses a love relation between the characters and currently 12 other social relations (see Appendix A for the complete list).

We suspect there would be no problems using these guidelines for fairy tales from different author or in different languages (one would probably need to expand the set of relation types as we expect to do when we annotate the remaining Grimm fairy tales). Outside of the domain of fairy tales, the character types of Propp will most likely not make much sense, so they would have to be discarded or replaced.

3.2 Annotation Procedure

For annotation, two annotators were asked to label all these layers on their own and afterwards had the task to discuss the differences and unify their individual solutions to form a final document. We measured inter-rater agreement of the individual layers using Cohen's Kappa: The values for age (61.2%), where the main source of disagreement was between the labels "unknown" and "aged", characters in the network (81.5%), entity sentiment (69.9%), character types (69.7%), sentiment between characters (71.5%) and social relations between characters (67.3% overall; 83.6% when only regarding family relations and 60.3% for all other relations) correspond to substantial agreement according to Landis and Koch (1977).

3.3 Dataset Statistics

The 40 annotated texts have an average length of 1823 tokens and contain a total of 11873 mentions in 1218 clusters (about 9.75 mentions per cluster). 589 of the clusters are singletons.

The networks contain 244 characters (6.1 plot-relevant characters per network). For detailed information about the distribution of the labels in the text files and in the networks, see Appendix A.

4 Baseline Experiments

In order to demonstrate the usefulness of our data set, both for the automatic extraction and evaluation of character networks, we conduct baseline experiments using a rule-based pipeline and an approach based on neural networks. Both approaches detect and cluster character references. The rule-based version uses the Named Entity Recognition of Jannidis et al. (2017) and the coreference module of Krug et al. (2015), while the neural network **c2f** (Lee et al., 2018) conducts both steps in an end-to-end fashion. c2f was pre-trained on DROC and fine-tuned on the fairy tales using 5 folds, so that the entirety of our corpus could be labeled by the system.[6] The resulting entities form the nodes in the networks. As a proxy for interactions (which represent the edges in the networks) we used co-occurence counts for a given discours (in our work we differentiate between a sentence or a paragraph as discours). The corresponding edge weights are the number of times the two entities appear in the same discours. Currently, we do not infer relation types, which is left for future work.

[6] We actually used an additional six fairy tales, for which we did not have character networks yet, for fine-tuning.

4.1 Social Network Analysis

The resulting networks are evaluated by two means. The first evaluation setting is directed at the network structure itself. We compare networks based on the gold coreference annotations and the system annotations using a set of metrics from Social Network Analysis:

- Average Weighted Degree: average number of neighbours of an entity weighted by the corresponding edge weight. A high degree means than an entity has interacted with a lot of other entities.
- Average Path Length: average length of all shortest paths in the graph.
- Network Diameter: longest of all shortest paths in the network.
- Graph Density: ratio of edges in the graph compared to the number of possible edges. The higher this number is, the more other entities each entity interacts with.
- Average Betweenness Centrality: average number of shortest paths that go through the node but do not start or end there. This is high if an entity connects many other entities to each other.

Table 1 shows the results of this experiment. No approach is currently capable of recovering the network structure (at least according to these statistics). For the discourse of sentences, the entities in the neural approach seem to be connected more strongly than required (as can be seen by the much higher density and the higher average degree) and the rule-based pipeline proposes entities that are linked too sparsely. Interestingly, this phenomenon is almost shifted when the discourse is changed to paragraphs, where the rule-based system appears to overestimate the inter-connectedness of the characters and the neural system underestimates it. Since these metrics do not compare a system and gold network directly and are difficult to interpret we performed a second experiment where the system networks were scored against the gold networks.

	Deg	PL	Dia	Den	B Cen
Gold Sent	15.69	2.80	6.05	0.33	18.92
Rules Sent	10.81	2.55	5.53	0.21	29.86
c2f Sent	19.56	2.82	5.70	0.53	8.86
Gold Para	13.22	1.93	3.38	0.48	7.33
Rules Para	15.75	1.58	2.77	0.40	11.47
c2f Para	11.77	1.70	3.00	0.64	3.66

Table 1: Average Weighted **Deg**ree, **P**ath **L**ength, **Dia**meter, **Den**sity and **B**etweenness **Cen**trality of entity graphs from the rule-based pipeline and c2f.

4.2 Precision at k

The second evaluation rates the quality of the most important entities that appear in the networks. For this setting, we extract a ranking of the k most important entities (using frequency of appearance in the text, which was shown by Krug et al. (2016) to correlate with the actual importance). The resulting rankings were matched in several ways against the gold entities:

- **Soft:** In this setting a suggested entity by the system is correct if it does appear at all in the network (no matter the rank)
- **Soft-Ranked:** In this setting the rank of the system and gold solution need to agree in order to obtain a correct match
- **Spearman-Ranked:** In this setting the entire ranking was compared to the ranking obtained from the gold networks using Spearman correlation Spearman (1904)

In order to find a corresponding system entity for each gold entity (which are both represented by clusters of mentions in the text), we use the Kuhn-Munkres algorithm used in the CEAF metric (Luo, 2005). The similarity of two entities is measured by the relative number of common mentions, like in the entity-based variant of CEAF. That means even if multiple system entities would be determined as compatible, only the "best" match will be chosen.

We introduced a cutoff parameter k in order to artificially limit the rankings at position k and therefore obtain insight into the quality at different levels of the ranking (e.g. is the ranking of the top 2 entities more reliable than that of the top 6).
Figure 1 shows Precision-at-k for soft (Soft) and soft-ranked gold entities as well as Spearman correlation for the results of the rule-based pipeline and the neural algorithm.

Interpreting the results first shows that the neural system outperformed the rule-based pipeline in every setting. In the **Soft** evaluation setting, the proposed most important entity is always part of

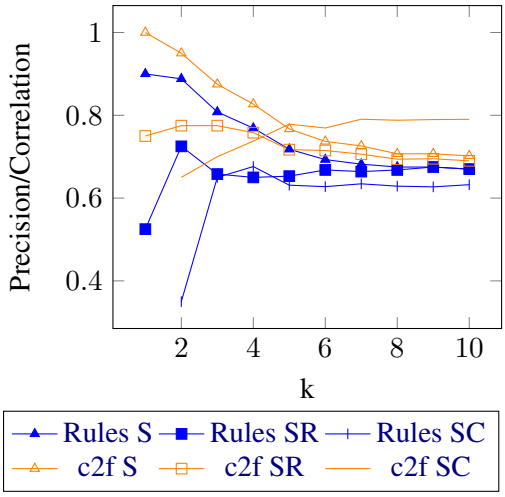

Figure 1: Precision and Spearman Correlation (SC) of the top k entities found via rule-based pipeline and neural algorithm (c2f). SR is short for the *Soft-ranked* setting and S is an abbreviation for the *Soft* setting

the network, but only about 75% of the time also the most important entity in the gold data. Both scores severly decline up to about 70% Precision with k at ten. That means, if we were to form a network comprising ten characters, three characters would be spurious and unwanted. In the most severe setting Spearman-Ranked, the correlation starts at about 0.75 (for k=2, since rankings of size 1 cannot be evaluated) and dropping to about 0.6 in the process. This correlation shows, that at least the relative importances of the entities seems to be maintained when an automatic attempt is made.

5 Conclusion

We present a corpus of 40 German fairy tales by the Brothers Grimm with hand-annotated coreference information and character networks. This data set can be used to train and compare different approaches for the automatic extraction of character networks and helps to homogenize the evaluation of research on this topic. We demonstrated its usefulness by comparing two different systems available for the automatic extraction of character networks. We found that even state-of-the-art approaches can not reliably extract character networks, even for a seemingly easy domain such as fairy tales, evaluated on the overlap of characters in the network and the general shape of the resulting graph. For future work, we plan to conduct experiments for the automatic prediction of all further annotated layers, especially sentiment, attributes and relations of/between the characters.

References

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada. Association for Computational Linguistics.

Mariona Coll Ardanuy and Caroline Sporleder. 2014. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 31–39, Gothenburg, Sweden. Association for Computational Linguistics.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Fotis Jannidis, Isabella Reger, Lukas Weimer, Markus Krug, Martin Toepfer, and Frank Puppe. 2017. Automatische Erkennung von Figuren in deutschsprachigen Romanen. *Proceedings of DhD*.

Prashant Jayannavar, Apoorv Agarwal, Melody Ju, and Owen Rambow. 2015. Validating literary theories using automatic social network extraction. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 32–41, Denver, Colorado, USA. Association for Computational Linguistics.

Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2007. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1040–1047, Prague, Czech Republic. Association for Computational Linguistics.

Bryan Jurish. 2012. *Finite-State Canonicalization Techniques for Historical German*. Ph.D. thesis, Universität Potsdam. (completed 2011, published 2012).

Markus Krug. 2020. *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. doctoralthesis, Universität Würzburg.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104, Denver, Colorado, USA. Association for Computational Linguistics.

Markus Krug, Frank Puppe, Fotis Jannidis, Isabella Reger, Lukas Weimer, and Luisa Macharowsky. 2016. Comparison of methods for the identification of main characters in german novels. In *Digital Humanities 2016: Conference Abstracts*, pages 578–582.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus, and Fotis Jannidis. 2018. Description of a corpus of character references in german novels - DROC [Deutsches ROman Corpus]. In *DARIAH-DE Working Papers*. DARIAH-DE.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Gyeong-Mi Park, Sung-Hwan Kim, Hye-Ryeon Hwang, and Hwan-Gue Cho. 2013. Complex system analysis of social networks extracted from literary fictions. *International Journal of Machine Learning and Computing*, 3(1):107.

Vladimir Propp. 1972. Morphologie des märchens [1928]. *Hrsg. Karl Eimermacher. Übs. Christel Wendt. München: Carl Hanser Verlag*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

A Inter Annotator Agreement and Corpus Statistics

Label	Precision	Recall	F1
unlabeled	96.9%	98.7%	97.7%
human	91.1%	95.7%	92.8%
legendary	83.1%	83.1%	79.0%
generic	85.8%	68.0%	73.1%
transcendent	95.1%	98.2%	96.2%
mixed	93.3%	91.7%	91.5%

Table 2: Agreement on annotated mention spans.

Metric	Precision	Recall	F1
MUC	96.5%	96.8%	96.7%
B^3	92.4%	91.0%	91.6%
$CEAF_E$	88.1%	84.3%	85.8%
$CEAF_M$	94.2%	92.5%	93.3%
LEA	90.0%	89.2%	89.9%

Table 3: Agreement on annotated mentions (including coreference id).

Label Type	Cohen's Kappa
Character Age	61.2
Character Type	69.7
Character Sentiment	69.9
Relation Type	67.3
Relation Sentiment	71.5

Table 4: Agreement on labels in the networks.

(a) Sex

Label	Ratio
female	29.0%
male	58.1%
unknown	12.9%

(b) Number

Label	Ratio
Singular	83.5%
Plural	13.0%
unknown	3.5%

(c) Syntactic Category

Label	Ratio
name	4.4%
noun phrase	30.8%
pronoun	64.8%

(d) Semantic Category

Label	Ratio
human	69.7%
legendary	24.0%
generic	5.8%
transcendent	0.4%
mixed	0.1%

Table 5: Distribution of Sex, Number, Syntactic Category and Semantic Category labels in the annotated texts.

(a) Sex

Label	Ratio
female	28.3%
male	56.6%
unknown	15.2%

(b) Sentiment

Label	Ratio
positive	31.1%
negative	19.7%
neutral	49.6%

(c) Character Type

Label	Ratio
dispatcher	10.2%
donor	3.3%
false hero	2.0%
helper	15.6%
hero	20.1%
plot relevant	32.8%
prize	4.9%
villain	13.1%

(d) Age

Label	Ratio
baby	0.8%
growing-up	16.0%
grown-up	32.8%
aged	7.0%
unknown	43.4%

Table 6: Distribution of Sex, Character Sentiment, Character Type and Age labels in the annotated character networks. Note that the percentages do not always add up to 100% since some entities can have several labels.

(a) Relation Type

Label	Ratio
none	57.0%
victim	4.2%
friend	9.9%
enemy	5.8%
desires	1.2%
loves	0.7%
parent	3.8%
child	3.5%
sibling	2.4%
spouse	3.6%
grandparent	0.3%
grandchild	0.3%
uncle	0.1%
nephew	0.1%
parent-in-law	1.1%
child-in-law	1.3%
stepparent	0.3%
stepchild	0.3%
stepsibling	0.1%
fosterparent	0.2%
fosterchild	0.2%
fostersibling	0.1%
ruler	1.8%
subject	1.4%
employee	1.3%
employer	1.3%
owner	0.6%
property	0.6%
tradepartner	0.6%
workmate	0.1%

(b) Relation Sentiment

Label	Ratio
none	51.7%
positive	16.0%
negative	11.5%
neutral	21.2%

Table 7: Distribution of Relation Type and Relation Sentiment labels in the annotated character networks. Note that the percentages do not always add up to 100% since some relations can have several labels.