Automatic detection and classification of literary character properties in German narratives

Janis Pagel[1], and Nils Reiter[1]

[1] Department for Digital Humanities, University of Cologne, Cologne, Germany

Abstract

This work presents an approach to automatically (i) identify sentences in German narrative texts that contain properties of literary characters and (ii) assign categories to these sentences according to what kind of property is described, both with coarse – such as role, clothing or physiognomy – and fine-grained categories – such as occupation, accessories or face. To this end, we test different transformer-based models (BERT, ELECTRA, RoBERTa, Llama) and compare the results to simple baselines (majority, random, bag-of-words Naive Bayes). We find that an uncased ELECTRA model achieves promising results in identifying sentences that contain character properties (67% F1), while uncased BERT achieves highest results in assigning coarse-grained categories to sentences (87% F1) and RoBERTa is the best model in assigning fine-grained categories (80% F1). A LoRA-tuned Llama 3.1 large language model is able to achieve comparable scores to the best encoder model on the coarse-grained task (81% F1), but is still 6 percentage points below the fine-tuned German BERT model.

Keywords: computational literary studies, literary character properties, transformers, large language models

1 Introduction

This paper presents a method to automatically identify snippets of German narrative texts in which a character property is mentioned and to classify the snippet according to the category of the property. Characters are crucial elements of narrative texts and automatically assigning properties to them is an important part of general, all-encompassing narrative understanding. At the same time, extracting character properties is a challenging task, as virtually any linguistic form can be used to assign a property to a character.

(1) Now Mr. Bumble was a fat man, and a choleric one […].[1]

(2) But nature or inheritance had implanted a good sturdy spirit in Oliver’s breast […].

(3) “What is it?” inquired the beadle.

(4) “Well, you have come here to be educated, and taught a useful trade,” said the red-faced gentleman in the high chair.

As an illustration, consider examples (1) to (4). Each sentence assigns one or more properties to a character through different linguistic means. In (1), the assignment is very explicit – a character with a proper name is described as being fat and choleric. Example (2) assigns a property –

Janis Pagel, and Nils Reiter. “Automatic detection and classification of literary character properties in German narratives.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1449–1464. https://doi.org/10.63744/UkAh6gmT12av.

[1] Examples (1) to (4) are taken from Charles Dicken’s *Oliver Twist*, Chapter 2.

having inner strength – through a description of the circumstances of his uprising. In (3), the character making the utterance is assigned the profession "beadle". Human readers have no trouble in determining that this is the character that previously was named Mr. Bumble. Finally, example (4) shows the use of an attributive adjective to describe the face on an unnamed gentleman. This list is by no means complete, but it illustrates the versatility of character property assignments in literary texts.

From a technical standpoint, a machine-readable representation of such a character assignment looks straightforward at first sight and can be expressed as a triple that links a character (id) through a property with a property value. For example, (1) could be expressed as `(mr-bumble, body_type, fat)`. In the practice of literary narratives, however, things are more complicated: Many property values can change, even throughout a single text. Thus, each triple needs to be associated with a timestamp information that encodes when within the narrated time the property assignment holds. Another complication arises from the fact that a description of a character in a narrative may be conducted with limited authority: Such descriptions can come from other characters (e.g., through direct or indirect speech) or appear in different narrative levels. In cases of unreliable narration or focalization [4], even the narrator may not tell us the whole story. Thus, in addition to temporal information, we would need to include information about the source of a property assignment.

There are different possibilities to deal with these complications. The solution we work towards is a fine-grained and modular approach: Assigning the source of a property assignment, determining its temporal position within the narrative world, determining the exact target of a property are all distinct sub-steps in a modular framework. This paper focuses on the first step: The detection of character property mentions in narrative prose.

2 Related Work

Characters are a prime component of narrative and dramatic literary texts and serve as anchor points for our perceived pleasure and/or identification when reading them. Literary characters have been studied extensively in literary studies, often with a focus on specific instances.[2] More systematically, there are publications about what exactly literary characters are [6; 18], how they can be grouped/classified/typed [8; 20] or what their significance for the text as a whole is [24]. There is also quite a bit of work on the comparison of multiple characters, within or across a specific text (e.g., [19] on Captain Ahab from Melville's *Moby Dick* and the ivory trader Kurtz from Conrad's *Heart of Darkness*). In sum, both individual literary characters are practically and the concept of literary character is theoretically well researched.

From this perspective, it may be somewhat surprising that they are not front and center in digital or computational approaches to literature. Due to the clear and fundamental differences in operationalization chances, the analysis of dramatic and prose characters rests on different assumptions and pre-conditions. Dramatic characters have been researched intensively in terms of their relations: Co-presence on stage with other characters (often in the form of social networks; [25; 26; 27]), family relations among the characters [30], or the knowledge they express about each other [1]. Besides work on identifying gender, age and social status [17], non-relational properties of dramatic characters have not been investigated using computational means. On prose texts, there is existing work on associating characters with their speech [5; 7], their sounds [11] and their emotions [15]. To our knowledge, only a small subset of character properties/aspects have been looked at so far in a generic way (i.e., not in a relation to a specific text or author), namely gender [23], their psychology [21] and the contents of character speech [12; 22; 28].

[2] For instance: [2] collects essays on Hermione Granger from Harry Potter, discussing her feminist nature; [29] discusses Sméagol/Gollum from Lord of the Rings.

There are few publications on the automatic identification/extraction of character properties in narratives in a generic way. [3] describe a system based on automatic syntactic and semantic analysis. Using a syntactic and semantic parser, the authors extract semantic roles for each character in Tolstois' *War and Peace* (in Russian) and with the help of principled component analysis find that some characters are more closely associated with Object, Agent, Addressee and other roles than others. [16] extract physical descriptions from Dutch-language "chick lit". Based on manually annotated sentences, they compare a machine learning with a lexical pattern-based approach and find the extraction via lexical patterns to achieve higher performance.

To our knowledge, we are the first to attempt a supervised extraction of theoretically motivated character properties in German narrative texts.

3 Automatic Detection of Character Properties

Our approach on the automatic extraction of character properties runs in multiple stages and this paper concentrates on the first one. The goal of this first stage is to i) identify spans in which one of a number of pre-defined character properties are mentioned and ii) classify the span according to the property. Thus, in a sentence like (5), the first stage system determines that this sentence mentions (among other things) the property "face".

(5) Er blickte scharf nach dem holden Angesicht, das sich einst im Zorn über ihn gerötet hatte.
He looked sharply at the fair face that had once reddened in anger at him.[3]

There are additional steps in order to conduct an end-to-end automatic extraction of fine-grained character properties: i) Extracting the property value (in (5), the fact that the face is described as "fair"), ii) identifying the character that this property (value) is attributed to. Both tasks are not the focus of this paper.

The concrete set of properties we are working with (shown in Table 1) has been developed in collaboration with project partners from literary studies. As a first step, a small set of texts has been annotated without pre-defined categories. Based on the annotated spans that convey a character property, we have defined a set of coarse and fine-grained property categories. Thus, they are not based on an ontological, systematic understanding of potential descriptions of humans, but are based on experiences and expectations of character descriptions that actually appear in (German-language) literary texts from a given time period. This leads to certain imbalances, as, for instance, the face is part of the head. Still, descriptions of faces or face aspects appear in high frequency and bear a high significance, that we decided to let them form their own category.

4 Data

The text sources for the annotated data comes from 14 texts found in d-prose [10] and three texts in TextGrid.[4,5] The texts were split into sentences using Spacy's sentence splitter[6] and manually annotated for the coarse- and fine-grained categories shown in Table 1.[7] The annotations were carried out by four trained German-speaking students of German literary studies. Next to the category, the annotators also marked which token span belongs to which category and which literary character is being described. Annotation guidelines (in German language) are used to detail the

[3] Translation by DeepL.

[4] `https://textgridrep.org/`

[5] A list of all texts is provided in Table 9, Appendix A.

[6] `https://spacy.io/api/dependencyparser`. Sentence splitting is based on the output of Spacy's dependency parser, see [13]. Model used: `de_core_news_sm`.

[7] An example for each fine-grained category in both the original German and an English translation is provided in Appendix B.

criteria on each coarse- and fine-grained category. Annotators are regularly supervised. Quality of the annotations is regularly checked through calculation of inter-annotator agreement.

Categories		Count
Coarse	Fine	
Age	Numerical	50
	Role w/ connection to age	235
	Scalar	201
	Total	486
Traits	Mind/habitus	28
	Basic attitude	17
	Body/health	19
	Standard of living	10
	Total	77
Clothing	Accessories	6
	Fashion appearance	1
	Piece of clothing	6
	Part of piece of clothing	1
	Total	14

Categories		Count
Coarse	Fine	
Physiognomy	Charisma	15
	Face	41
	Finger/hand/arm	2
	Head/hair	24
	Height/stature/weight	45
	Trunk/shoulder	4
	Toe/foot/leg	10
	Total	141
Role	Occupation	148
	Relationship	45
	Sex	370
	Family	462
	Nationality/place of origin	14
	Religion/politics	7
	Type	67
	Social status	68
	Total	1186

Table 1: Coarse- and fine-grained categories as annotated. Each annotation is attached to a single sentence.

Inter-annotator agreement was measured on three of the texts mentioned in Table 9.[8] Overall, there is an agreement among all four annotators, measured in Fleiss κ [9], of 0.23 for the fine-grained categories and 0.93 for the coarse-grained categories, suggesting that the later task is much easier for humans to perform than the former. When looking at single fine-grained categories, the ones with the lowest agreement were *type* and *part of piece of clothing* with a κ value of -0.001 and the ones with the highest agreement were *finger/hand/arm* and *social status* with κ values of 0.499 and 0.497, respectively.[9] Note that some of these fine-grained categories appear with extremely low frequency, which also affects the measurement of inter-annotator agreement. It should further be noted that even for the fine-grained categories with the highest agreements, scores of around 0.5 are usually not considered to signify high agreement in general. In subsequent model training for the three texts used to determine inter-annotator agreement, only annotations by one annotator have been used.

5 Experimental Setup

We perform 5-fold cross-validation with five repetitions, meaning that the folds are additionally randomly split five times and the result is averaged. Furthermore, we make sure that all classes have the same ratio per train/test set and fold like in the complete dataset. This results in the distribution of categories in Table 2, averaged over all folds and repetitions. Since for the binary task, there would be a large imbalance of sentences containing and not containing a character property, we apply downsampling to the number of sentences that do contain a character property in the train

[8] These three texts are “Der Selbstmordverein”, “Altmodische Leute” and “Der Katzenjunker”.

[9] A full list of agreement scores for the coarse-grained and fine-grained categories can be found in Appendix C.

set, while keeping the test set intact. For each sentence in the training and test set, we prepend the two previous and append the two following sentences as context for the models.

For automatic prediction, we use the following encoder models: A cased and uncased version of German BERT[10], a cased and uncased version of German ELECTRA[11] and German RoBERTa[12]. We fine-tune all models for 20 epochs, using a learning rate of 4×10^{-5}. Additionally, we compare the results of the BERT-like models for the coarse-grained category task to one decoder large language model, namely Llama-3.1 8B Instruct[13], by fine-tuning the model on one fold of the training data, using LoRA [14], for 10 epochs, a learning rate of 1×10^{-4} and a rank of 32. For both LoRA finetuning and prediction, the prompt in Listing 1 was used.

```
Give a character property label to the following text snippet!

        The following labels are possible:
        - Age
        - Character trait
        - Clothing
        - Physiognomy
        - Role

Do not output anything else!!!
```

Listing 1: Prompt to fine-tune the LLM with LoRA.

For the binary task, we also used the predictions of a named entity recognition model[14] on the sentences and added a one-hot encoded vector to the input embeddings of all models, containing the information if the (sub-)token is labeled as a named entity or not.

For the coarse- and fine-grained tasks, we run two lines of experiments: for one line, we add XML tags into the input string that tell the models where the target sentence starts and ends. This annotation provides the models information on where to divide between actual sentence in question and mere context. For the other line, these tags are missing. This way we can test if providing this information is helpful for the transformer models to make decisions. For the binary task where the model is supposed to detect if a character property is present in the sentence or not, we never provide the sentence markers, as this information would already preempt the answer and would give the models too much information.

It is also important to note that for the coarse- and fine-grained tasks, the model has access to the sentences that contain a character property according to our annotators, so the tasks show the upper bound of possible classification scores for character property classes.

We also compare the results of the transformer models to some simple baselines: a majority baseline (most frequent), two random baselines, (i) were the classes are randomly picked according to the frequency of the class in the training data (Random (stratified)), (ii) were the classes are picked entirely randomly (Random (uniform)) and a Naive Bayes model with features based on the counts of a bag-of-words transformation of the training sentences.

[10] `https://huggingface.co/dbmdz/bert-base-german-uncased`, `https://huggingface.co/bert-base-german-cased`

[11] `https://huggingface.co/german-nlp-group/electra-base-german-uncased`, `dbmdz/electra-base-german-europeana-cased-discriminator`

[12] `https://huggingface.co/benjamin/roberta-base-wechsel-german`

[13] `meta-llama/Llama-3.1-8B-Instruct`

[14] `https://huggingface.co/flair/ner-german`

Category	Train		Test	
	mean	sd	mean	sd
No character property	1523.2	0.4	3873.0	0.0
Character property present	1523.2	0.4	380.8	0.4
Age	388.8	0.4	97.2	0.4
Character trait	61.6	0.5	15.4	0.5
Clothing	11.2	0.4	2.8	0.4
Physiognomy	112.8	0.4	28.2	0.4
Role	948.8	0.4	237.2	0.4
Accessories	4.8	0.4	1.2	0.4
Basic attitude	13.6	0.5	3.4	0.5
Body/health	15.2	0.4	3.8	0.4
Charisma	12.0	0.0	3.0	0.0
Face	32.8	0.4	8.2	0.4
Family	369.6	0.5	92.4	0.5
Head/hair	19.2	0.4	4.8	0.4
Height/stature/weight	36.0	0.0	9.0	0.0
Mind/habitus	22.4	0.5	5.6	0.5
Nationality/place of origin	11.2	0.4	2.8	0.4
Numerical age	40.0	0.0	10.0	0.0
Occupation	118.4	0.5	29.6	0.5
Piece of clothing	4.8	0.4	1.2	0.4
Relationship	36.0	0.0	9.0	0.0
Religion/politics	5.6	0.5	1.4	0.5
Role with connection to age	188.0	0.0	47.0	0.0
Scalar age	160.8	0.4	40.2	0.4
Sex	296.0	0.0	74.0	0.0
Social status	54.4	0.5	13.6	0.0
Standard of living	8.0	0.0	2.0	0.0
Toe/foot/leg	8.0	0.0	2.0	0.0
Trunk/shoulder	3.2	0.4	1.0	0.0
Type	53.6	0.5	13.4	0.5

Table 2: Distribution of categories across the train and test set, averaged for each fold and repetition. In total, there are less instances for the fine-grained categories (378) in the test set than for the coarse-grained categories (381), because some fine-grained categories are excluded for having too less instances (part of piece of clothing, fashion appearance, finger/hand/arm).

6 Results

As previously mentioned, we evaluate classification performance on three tasks of increasing label complexity: a binary classification (detecting whether a sentence contains a character property), a 5-class coarse-grained classification and a fine-grained 18-class classification into specific subcategories. We report macro-averaged F1, precision, recall and accuracy for each model. While micro-averaged scores would be higher than the macro-averaged variants, we are mainly interested in the overall performance of the models independent of the contributions of each category.

Model	F1		Precision		Recall		Accuracy	
	mean	sd	mean	sd	mean	sd	mean	sd
Most frequent	0.65	0.00	**0.91**	0.00	0.50	0.00	**0.91**	0.00
Random (stratified)	0.50	0.01	0.50	0.00	0.50	0.01	0.50	0.00
Random (uniform)	0.50	0.01	0.50	0.00	0.50	0.01	0.50	0.00
Naive Bayes	0.66	0.01	0.59	0.00	0.74	0.01	0.72	0.01
BERT (uncased)	0.58	0.28	0.57	0.32	0.71	0.20	0.66	0.36
BERT (cased)	0.52	0.29	0.53	0.35	0.66	0.20	0.59	0.39
ELECTRA (uncased)	**0.67**	0.21	0.67	0.24	**0.75**	0.20	0.76	0.28
ELECTRA (cased)	0.55	0.23	0.65	0.34	0.59	0.16	0.69	0.35
RoBERTa	0.48	0.32	0.55	0.42	0.58	0.18	0.53	0.43

Table 3: Overall results on binary character property detection task. Metrics are macro-averaged. *Mean* shows the average over all 5 folds with 5 repetitions, *sd* the standard deviation. Bold indicates the best result in each column. The upper part shows baseline performances.

As shown in Tables 3–5, the transformer-based models outperform the baseline classifiers on all tasks; only for the binary task, the majority baseline has the highest precision. On the binary character property detection, the best model (uncased ELECTRA) achieves an averaged macro-F1 of 0.67. It is notable that the Naive Bayes baseline almost reaches the performance of the ELECTRA model with just one percentage point of difference (0.66), suggesting that lexical clues are enough to detect many character properties and that the transformer models are able to correctly classify these instances as well, but are not able to pick up on more implicit features of the sentences where this is not the case. Furthermore, we observe very high standard deviations for the transformer model results across the folds and repetitions, suggesting that the dataset is rather diverse and that it matters which dataset splits to present to the models. For the 5-class coarse-grained categories, the top model reaches 0.87 F1 (uncased BERT), far exceeding the uniform baseline's 0.18. Also, adding sentence markers to tell the model which sentence to focus on often helps with prediction, albeit to a varying degree. Especially the cased ELECTRA model seems to be thrown off by the sentence markers, as adding a sentence marker reduces this model's performance by 16–34 percentage points. We also notice that using the uncased version of a model over the cased version generally improves the results significantly. Note that many models show a certain discrepancy between accuracy and F1 score, which can be explained by the fact that F1 is calculated as the macro-average. The fine-grained 18-class task is slightly more challenging: the highest macro-F1 is 0.8 (with RoBERTa), though this still represents a large improvement over the near-zero majority (0.07) and random baselines (0.05 and 0.04). We do not observe large differences between cased and uncased versions of models.

When adding one-hot encoded embeddings containing information about the named entity classes of a sentence to the models, as described in Section 5, we do not find any difference in performance for any of the models.

		F1		Precision		Recall		Accuracy	
Target	Model	mean	sd	mean	sd	mean	sd	mean	sd
Unmarked	Most frequent	0.30	0.00	0.62	0.00	0.20	0.00	0.62	0.00
	Random (stratified)	0.21	0.01	0.21	0.01	0.20	0.01	0.48	0.01
	Random (uniform)	0.18	0.03	0.19	0.01	0.17	0.05	0.18	0.01
	Naive Bayes	0.31	0.02	0.70	0.13	0.20	0.00	0.63	0.00
	BERT (uncased)	0.83	0.17	0.89	0.10	0.80	0.21	0.94	0.08
	BERT (cased)	0.74	0.26	0.84	0.13	0.69	0.31	0.88	0.15
	ELECTRA (uncased)	0.86	0.11	0.90	0.06	0.83	0.15	**0.95**	0.03
	ELECTRA (cased)	0.84	0.14	0.90	0.05	0.81	0.19	0.95	0.04
	RoBERTa	0.79	0.23	0.90	0.11	0.74	0.28	0.92	0.12
Marked	Most frequent	0.30	0.00	0.62	0.00	0.20	0.00	0.62	0.00
	Random (stratified)	0.21	0.01	0.21	0.01	0.20	0.01	0.48	0.01
	Random (uniform)	0.18	0.03	0.19	0.01	0.17	0.05	0.18	0.01
	Naive Bayes	0.31	0.02	0.68	0.13	0.20	0.00	0.62	0.00
	BERT (uncased)	**0.87**	0.15	**0.91**	0.07	**0.85**	0.19	**0.95**	0.07
	BERT (cased)	0.75	0.24	0.86	0.12	0.71	0.29	0.90	0.13
	ELECTRA (uncased)	0.84	0.14	0.85	0.13	0.83	0.15	0.92	0.10
	ELECTRA (cased)	0.54	0.30	0.74	0.15	0.47	0.35	0.76	0.17
	RoBERTa	0.80	0.27	0.92	0.09	0.76	0.33	0.90	0.15

Table 4: Overall, macro-averaged results on the coarse-grained classification task (5 classes). *Mean* shows the average over all 5 folds with 5 repetitions, *sd* the standard deviation. Bold font marks the top value per column.

		F1		Precision		Recall		Accuracy	
Target	Model	mean	sd	mean	sd	mean	sd	mean	sd
Unmarked	Most frequent	0.07	0.00	0.24	0.00	0.04	0.00	0.24	0.00
	Random (stratified)	0.05	0.01	0.05	0.01	0.05	0.01	0.14	0.02
	Random (uniform)	0.04	0.01	0.04	0.01	0.04	0.02	0.04	0.01
	Naive Bayes	0.09	0.00	0.46	0.11	0.05	0.00	0.27	0.01
	BERT (uncased)	0.74	0.11	0.75	0.10	0.73	0.11	0.75	0.07
	BERT (cased)	0.74	0.11	0.75	0.11	0.73	0.11	0.75	0.06
	ELECTRA (uncased)	0.73	0.12	0.74	0.12	0.72	0.13	0.75	0.07
	ELECTRA (cased)	0.71	0.15	0.73	0.14	0.69	0.16	0.75	0.09
	RoBERTa	**0.80**	0.06	**0.81**	0.06	**0.80**	0.06	**0.79**	0.03
Marked	Most frequent	0.07	0.00	0.24	0.00	0.04	0.00	0.24	0.00
	Random (stratified)	0.05	0.01	0.05	0.01	0.05	0.01	0.14	0.02
	Random (uniform)	0.04	0.01	0.04	0.01	0.04	0.02	0.04	0.01
	Naive Bayes	0.09	0.00	0.46	0.12	0.05	0.00	0.27	0.01
	BERT (uncased)	0.76	0.10	0.77	0.10	0.75	0.10	0.76	0.05
	BERT (cased)	0.75	0.10	0.76	0.10	0.74	0.10	0.76	0.05
	ELECTRA (uncased)	0.74	0.12	0.75	0.11	0.73	0.12	0.76	0.07
	ELECTRA (cased)	0.74	0.12	0.75	0.11	0.74	0.13	0.76	0.06
	RoBERTa	**0.80**	0.06	**0.81**	0.06	0.79	0.06	**0.79**	0.04

Table 5: Overall results on the fine-grained classification task (18 classes). *Mean* shows the average over all 5 folds with 5 repetitions, *sd* the standard deviation. Bold values indicate best-in-column results.

A closer per-class analysis of the fine-grained category predictions in Table 8 reveals the impact of label imbalance. The best models for each task perform reasonably well on the frequent classes but struggle on the rare ones. For example, the RoBERTa classifier attains $F_1 = 0.78$ on the most common fine-grained category *family* ($\sim$ 370 train instances, $\sim$ 92 test instances), but it fails to perform well for several infrequent categories – *religion/politics* ($\sim$ 6 train instances, $\sim$ 1 test instance) and *nationality/place of origin* ($\sim$ 11 train instances, $\sim$ 2 test instances) have $F_1 = 0.24$ and $F_1 = 0.41$ with or without sentence markers present. A similar imbalance effect is observed in the coarse-grained task dominated by the *role* class (Table 7).

	F1		Precision		Recall		Support
Class	mean	sd	mean	sd	mean	sd	mean
Character property present	0.38	0.26	0.29	0.20	0.73	0.42	380.8
No character property	0.80	0.32	0.85	0.32	0.77	0.32	3873.0

Table 6: Per-class results for binary classification task and the best model ELECTRA (uncased).

			BERT (uncased)						Llama 3.1 w/ LoRA		
		Support	F1		Precision		Recall		F1	Precision	Recall
Target	Category	mean	mean	sd	mean	sd	mean	sd			
Marked	Age	97.2	0.92	0.20	0.91	0.20	0.94	0.20	0.93	1.00	0.87
	Character trait	15.4	0.70	0.23	0.69	0.24	0.72	0.26	0.56	0.56	0.56
	Clothing	2.8	0.78	0.36	0.82	0.37	0.75	0.37	0.75	1.00	0.60
	Physiognomy	28.2	0.86	0.27	0.88	0.27	0.85	0.27	0.88	0.89	0.86
	Role	237.2	0.98	0.04	0.98	0.07	0.99	0.01	0.94	0.91	0.98
Unmarked	Age	97.2	0.92	0.19	0.90	0.20	0.94	0.20	0.93	0.99	0.87
	Character trait	15.4	0.59	0.32	0.60	0.31	0.59	0.33	0.39	0.38	0.40
	Clothing	2.8	0.70	0.40	0.75	0.41	0.69	0.41	0.50	0.67	0.40
	Physiognomy	28.2	0.79	0.33	0.83	0.32	0.77	0.34	0.87	0.96	0.79
	Role	237.2	0.97	0.05	0.96	0.08	0.99	0.01	0.93	0.89	0.97

Table 7: Per-class results for the coarse-grained category classification task, showing BERT (uncased) and Llama 3.1 w/ LoRA.

The results on the coarse-grained task for the LoRA-fine-tuned LLAMA model in Table 7 are fairly strong, ranging between 0.39 and 0.94 macro F1-score depending on the category and if the target was marked or not. As for the BERT-like models, *role* and *age* are predicted with the best performance scores. However, uncased BERT achieves higher F1 scores for all categories except *physiognomy* and *age*.

7 Conclusion

This paper introduces a comprehensive approach to detecting and categorizing character property mentions in German narrative prose. We showed that transformer-based models are generally able to perform three tasks of increasing complexity: binary classification (whether a character property is mentioned), coarse-grained category classification and fine-grained classification. One could suspect that some of the categories are strongly lexicalized, as it is, for instance, difficult to talk about character age without using very specific and unambiguous vocabulary (“old”, “young”, “years”, …). The strong performance of the bag-of-words-based Naive Bayes classifier for the binary task shows that this could be the case; however, this baseline did not perform as strongly for the coarse- and fine-grained category tasks.

		F1		Precision		Recall		Support
Target	Category	mean	sd	mean	sd	mean	sd	mean
Marked	Accessories	0.99	0.07	1.00	0.00	0.98	0.10	1.2
	Basic attitude	0.91	0.12	0.93	0.13	0.91	0.16	3.4
	Body/health	0.83	0.11	0.86	0.16	0.85	0.16	3.8
	Charisma	0.97	0.13	0.97	0.13	0.97	0.13	3.0
	Face	0.98	0.07	0.98	0.07	0.98	0.07	8.2
	Family	0.78	0.05	0.79	0.05	0.78	0.07	92.4
	Head/hair	0.99	0.07	0.99	0.05	0.98	0.08	4.8
	Height/stature/weight	0.85	0.07	0.88	0.13	0.85	0.14	9.0
	Mind/habitus	0.95	0.08	0.95	0.12	0.96	0.08	5.6
	Nationality/place of origin	0.41	0.22	0.47	0.32	0.43	0.28	2.8
	Numerical age	0.98	0.03	0.99	0.03	0.98	0.04	10.0
	Occupation	0.56	0.08	0.57	0.10	0.56	0.09	29.6
	Piece of clothing	1.00	0.00	1.00	0.00	1.00	0.00	1.2
	Relationship	0.54	0.21	0.56	0.19	0.56	0.27	9.0
	Religion/politics	0.24	0.34	0.23	0.34	0.32	0.43	1.4
	Role with connection to age	0.96	0.04	0.96	0.04	0.96	0.05	47.0
	Scalar age	0.92	0.03	0.93	0.04	0.92	0.04	40.2
	Sex	0.72	0.05	0.73	0.05	0.71	0.07	74.0
	Social status	0.67	0.11	0.65	0.13	0.69	0.14	13.6
	Standard of living	0.93	0.17	0.97	0.15	0.92	0.19	2.0
	Toe/foot/leg	0.98	0.10	0.98	0.10	0.98	0.10	2.0
	Trunk/shoulder	1.00	0.00	1.00	0.00	1.00	0.00	1.0
	Type	0.57	0.13	0.60	0.17	0.59	0.20	13.4
Unmarked	Accessories	0.99	0.07	1.00	0.00	0.98	0.10	1.2
	Basic attitude	0.90	0.11	0.92	0.11	0.91	0.16	3.4
	Body/health	0.83	0.11	0.85	0.17	0.85	0.15	3.8
	Charisma	0.98	0.12	0.98	0.10	0.97	0.13	3.0
	Face	0.99	0.05	0.99	0.07	1.00	0.03	8.2
	Family	0.78	0.04	0.79	0.05	0.78	0.07	92.4
	Head/hair	0.99	0.05	1.00	0.00	0.98	0.08	4.8
	Height/stature/weight	0.85	0.10	0.89	0.13	0.85	0.15	9.0
	Mind/habitus	0.94	0.09	0.95	0.10	0.95	0.10	5.6
	Nationality/place of origin	0.41	0.24	0.47	0.34	0.43	0.29	2.8
	Numerical age	0.98	0.02	0.99	0.03	0.98	0.04	10.0
	Occupation	0.57	0.08	0.57	0.08	0.57	0.10	29.6
	Piece of clothing	0.99	0.07	0.98	0.10	1.00	0.00	1.2
	Relationship	0.53	0.21	0.54	0.19	0.55	0.26	9.0
	Religion/politics	0.24	0.34	0.23	0.34	0.32	0.43	1.4
	Role with connection to age	0.96	0.03	0.97	0.03	0.97	0.05	47.0
	Scalar age	0.93	0.02	0.93	0.04	0.93	0.04	40.2
	Sex	0.72	0.05	0.72	0.05	0.71	0.08	74.0
	Social status	0.67	0.10	0.67	0.13	0.70	0.13	13.6
	Standard of living	0.99	0.07	1.00	0.00	0.98	0.10	2.0
	Toe/foot/leg	0.98	0.10	0.98	0.10	0.98	0.10	2.0
	Trunk/shoulder	1.00	0.00	1.00	0.00	1.00	0.00	1.0
	Type	0.57	0.12	0.62	0.19	0.57	0.17	13.4

Table 8: Per-class results for the fine-grained category classification task and the best model RoBERTa.

The most challenging task for the models is the detection of a character property in a sentence, where the best model ELECTRA (uncased) achieves an F1-score of 0.67.

We also showed that the inclusion of sentence markers yields systematic gains in classification performance across most models and tasks for the coarse-grained categories. This suggests that providing context in a structured form to the model can help mitigate ambiguity and improve the focus on relevant spans. By contrast, encoding named entity recognition (NER) tags as additional input features had negligible impact, indicating that entity-type information alone does not benefit character property recognition in this domain.

Our analysis also revealed that label imbalance remains a central challenge, especially in the fine-grained task. Frequent labels such as family or age were classified with reasonable success, while rare labels like nationality or religion/politics were often missed entirely. This effect is also observable for the coarse-grained task, although to a lesser extent.

The often high standard deviations across folds for all small transformer models raises the question of generalizability and especially the question of how heterogeneous the data actually is. Looking further into different properties of the sentence, e.g. linguistic features, might reveal that the models have no issue with certain types of sentences but struggle with others.

Notably, a fine-tuned LLaMA 3.1 model using parameter-efficient LoRA adaptation achieved competitive results on the coarse-grained task, rivaling or surpassing transformer encoders. This underscores the growing relevance of decoder-style large language models for classification tasks in computational literary studies, particularly when paired with lightweight tuning methods such as LoRA. However, we also observe LLM-related issues: In some cases during our experiments, the model invented new categories, i.e., hallucinated a new label. Any kind of LLM-based classification needs to be ready to deal with such issues (and count them as errors for evaluation purposes). A tempting alternative to using LLMs with LoRA adaptation is the prompting of a raw language model. Initial experiments revealed much weaker performance. In addition, hallucination issues might become more frequent.

In sum, this work lays a robust foundation for the automatic extraction of literary character properties in German-language prose, advancing computational literary studies' research on literary characters. Future work will extend this framework to model additional dimensions of character description, including temporal anchoring, source attribution and coreference resolution, moving toward a full-fledged representation of literary character descriptions.

Acknowledgments

We would like to thank our colleagues from the CompAnno project, Julia Nantke and Marie Flüh, for fruitful collaboration and discussions and the annotated data. We also thank the German Research Foundation (DFG) for funding the CompAnno project (project number 508319395). All experiments were carried out using the computational facilities of the Center for Data and Simulation Science (CDS) at the University of Cologne, for which we are also thankful.

References

[1] Andresen, Melanie, Krautter, Benjamin, Pagel, Janis, and Reiter, Nils. “Who Knows What in German Drama? A Composite Annotation Scheme for Knowledge Transfer”. In: *Journal of Computational Literary Studies* (2022). DOI: `10.48694/JCLS.107`.

[2] Christopher E. Bell, edited by. “Hermione Granger Saves the World: Essays on the Feminist Heroine of Hogwarts”. McFarland & Company Publishers, Inc., 2012.

[3] Bonch-Osmolovskaya, Anastasia and Skorinkin, Daniil. “Text mining War and Peace: Automatic extraction of character traits from literary pieces”. In: *Digital Scholarship in the Humanities* (Dec. 31, 2016), pp. i17–i24. DOI: `10.1093/llc/fqw052`.

[4] Bruhns, Adrian and Köppe, Tilmann. “Internal Focalization and Seeing through a Character’s Eyes”. In: *Estetika: The European Journal of Aesthetics 61*, no. 2 (Sept. 12, 2024). DOI: 10.33134/eeja.364.

[5] Brunner, Annelen, Tu, Ngoc Duyen Tanja, Weimer, Lukas, and Jannidis, Fotis. “To BERT or not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation”. In: *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. KONVENS. Online, 2020. URL: https://ceur-ws.org/Vol-2624/paper5.pdf.

[6] Eaton, Marcia M. “On Being A Character”. In: *The British Journal of Aesthetics 16*, no. 1 (Jan. 1976), pp. 24–31. DOI: 10.1093/bjaesthetics/16.1.24.

[7] Elson, David K. and McKeown, Kathleen. “Automatic Attribution of Quoted Speech in Literary Narrative”. In: *AAAI Conference on Artificial Intelligence*. 2010. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1945.

[8] Fishelov, David. “Types of Character, Characteristics of Types”. In: *Style 24*, no. 3 (1990), pp. 422–439.

[9] Fleiss, Joseph L. “Measuring Nominal Scale Agreement Among Many Raters”. In: *Psychological Bulletin 76*, no. 5 (1971), pp. 378–382.

[10] Gius, Evelyn, Guhr, Svenja, and Uglanova, Inna. ““d-Prose 1870–1920” a Collection of German Prose Texts from 1870 to 1920”. In: *Journal of Open Humanities Data 7*, no. 11 (July 2021), pp. 1–5. DOI: 10.5334/johd.30.

[11] Guhr, Svenja and Algee-Hewitt, Mark. “What’s that Scary Sound? Ambient Sound in Gothic Fiction”. In: *Journal of Computational Literary Studies* (Mar. 24, 2024). DOI: 10.48694/JCLS.3583.

[12] Hess, Leopold and Bary, Corien. “Narrator language and character language in Thucydides: A quantitative study of narrative perspective”. In: *Digital Scholarship in the Humanities* (June 25, 2019). DOI: 10.1093/llc/fqz026.

[13] Honnibal, Matthew and Johnson, Mark. “An Improved Non-monotonic Transition System for Dependency Parsing”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal, Sept. 2015, pp. 1373–1378. DOI: 10.18653/v1/D15-1162. URL: https://aclanthology.org/D15-1162/.

[14] Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, and Chen, Weizhu. “LoRA: Low-Rank Adaptation of Large Language Models”. Version 2. 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[15] Kim, Evgeny and Klinger, Roman. “Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions”. In: *Proceedings of the 27th International Conference on Computational Linguistics*, ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. ACL, Aug. 2018, pp. 1345–1359. URL: https://aclanthology.org/C18-1114.

[16] Koolen, Corina and Van Cranenburgh, Andreas. “Blue eyes and porcelain cheeks: Computational extraction of physical descriptions from Dutch chick lit and literary novels”. In: *Digital Scholarship in the Humanities 33*, no. 1 (Apr. 1, 2018), pp. 59–71. DOI: 10.1093/llc/fqx016.

[17] Krautter, Benjamin, Pagel, Janis, Reiter, Nils, and Willand, Marcus. “Properties of Dramatic Characters: Automatically Detecting Gender, Age, and Social Status”. In: *Computational Stylistics in Poetry, Prose, and Drama*. De Gruyter, Dec. 5, 2022, pp. 179–202. DOI: 10.1515/9783110781502-010.

[18] Margolin, Uri. “The What, the When, and the How of Being a Character in Literary Narrative”. In: *Style 24*, no. 3 (1990), pp. 453–468. URL: http://www.jstor.org/stable/42945873.

[19] Meyers. “The Fateful Impact: Moby-Dick and Heart of Darkness”. In: *Style 52*, no. 3 (2018), p. 212. DOI: 10.5325/style.52.3.0212.

[20] Propp, Vladimir Yakovlevich. *Morphology of the Folktale*. 2nd. Austin, TX: University of Texas Press, 1958.

[21] Rashkin, Hannah, Bosselut, Antoine, Sap, Maarten, Knight, Kevin, and Choi, Yejin. “Modeling Naive Psychology of Characters in Simple Commonsense Stories”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2289–2299. DOI: 10.18653/v1/P18-1213.

[22] Rybicki, Jan. “Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz’s Trilogy and its Two English Translations”. In: *Digital Scholarship in the Humanities 21*, no. 1 (Apr. 1, 2006), pp. 91–103. DOI: 10.1093/llc/fqh051.

[23] Schumacher, Mareike and Flüh, Marie. “StanfordNER Gender-Classifier”. Version 1.0. Oct. 8, 2021. DOI: 10.5281/ZENODO.5555952.

[24] Šušić, Mirela. “Methodical Approach to a Literary Character”. In: *European Journal of Language and Literature 6*, no. 2 (Oct. 15, 2020), p. 93. DOI: 10.26417/237rwf56t.

[25] Szemes, Botond and Vida, Bence. “Tragic and Comical Networks: Clustering Dramatic Genres According to Structural Properties”. In: *Computational Drama Analysis*. De Gruyter, June 17, 2024, pp. 167–188. DOI: 10.1515/9783111071824-009.

[26] Trilcke, Peer. “Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft”. In: *Empirie in der Literaturwissenschaft*, ed. by Philip Ajouri, Katja Mellmann, and Christoph Rauem. Münster, Germany, 2013, pp. 201–247.

[27] Trilcke, Peer, Fischer, Frank, and Kampkaspar, Dario. “Digital Network Analysis of Dramatic Texts”. In: *DH2015 Conference Abstracts*. 2015.

[28] Vishnubhotla, Krishnapriya, Hammond, Adam, and Hirst, Graeme. “Are Fictional Voices Distinguishable? Classifying Character Voices in Modern Drama”. In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, June 2019, pp. 29–34. URL: http://www.aclweb.org/anthology/W19-2504.

[29] Walch, Kathrin. “Sméagol – Individualpsychologische Analyse einer Romanfigur”. In: *Zeitschrift für freie psychoanalytische Forschung und Individualpsychologie* (May 1, 2015), 42–56 Seiten. DOI: 10.15136/2015.2.1.42-56.

[30] Wiedmer, Nathalie, Pagel, Janis, and Reiter, Nils. “Romeo, Freund des Mercutio: Semi-Automatische Extraktion von Beziehungen zwischen dramatischen Figuren”. In: *Abstracts of DHd*. Mar. 2020.

A Texts

Table 9 shows the 19 texts that were the bases for the annotations.

Title	Author	Publication Year
Der blonde Eckbert	Tieck, Ludwig	1797
Das Erdbeben in Chili	Kleist, Heinrich von	1807
Die Judenbuche	Droste-Hülshoff, Annette von	1842
Marcus König	Freytag, Gustav	1876
Der Katzenjunker	François, Louise von	1879
Krambambuli	Ebner-Eschenbach, Marie von	1883
Der Scout	May, Karl	1888
Altmodische Leute	Frapan, Ilse	1890
Die Schlangendame	Bierbaum, Otto Julius	1896
Die Frau Bürgermeisterin	Ebers, Georg	1897
Amazonenschlacht	Janitschek, Maria	1897
Kerlchen als Anstandsdame	Rose, Felicitas	1900
Münchhausen und Clarissa	Scheerbart, Paul	1906
Lena S.	Meyer Förder, Wilhelm	1908
Die Verwandlung	Kafka, Franz	1915
Der Selbstmordverein	Reventlow, Franziska Gräfin zu	1916
Das Liebesleben eines deutschen Jünglings	Zapp, Arthur	1920

Table 9: The 17 texts that were used as the basis for all analysis in this study.

B Examples for Categories

Table 10 provides a short example sentence for each fine-grained category. The examples are originally in German and additionally translated into English.

Category	Example	English Translation
Family	**Mein Vater** wird nächstens Geheimrat werden.	**My father** will soon become a privy councilor.
Sex	Dieser Brief hinterließ in **Herrn** Brock junior fatale Gefühle.	This letter left **Mr.** Brock junior with fatal feelings.
Role with connection to age	Das **Mädchen**: »Na, eigentlich heiß ich Mathilde.	The **girl**: ”Well, actually my name is Mathilde.
Scalar age	Der **alte** Pfadfinder schien ein ganz anderer Mensch geworden zu sein.	The **old** scout seemed to have become a completely different person.
Occupation	Das Mädchen sprach: »Nein, **Herr Doktor**!	The girl said: ”No, **Mr. Doctor**!
Social status	Ich grüße Euch mein Kumpan, **Herzog** Albrecht von Brandenburg!	I greet you, my comrade, **Duke** Albrecht of Brandenburg!
Type	»Paul, Du mußt mich nicht für ein **Nilpferd** halten; das ist beleidigend.«	“Paul, you mustn’t think of me as a **hippopotamus**; that’s offensive.”

Numerical age	»Ich denke: **So an die fünfundzwanzig**.	”I think: **About twenty-five**.
Relationship	Ich muß wohl **sehr verliebt in Dich sein**.	I must be **very much in love with you**.
Height/stature/weight	Resigniert **blickte er, so weit es ging, an seinem Bauch hinab**.	Resigned, he **looked down at his belly, as far as he could**.
Face	”Hat sie nicht unter blonden Haaren **braune Augen**?	”Doesn’t she have **brown eyes** under her blonde hair?
Mind/habitus	Noch war Gregor hier und **dachte nicht im geringsten daran, seine Familie zu verlassen**.	Gregor was still here and **did not think in the least of leaving his family**.
Head/hair	Er hatte eine **Platte**.	He was **bald**.
Body/health	Da sitzt er mit **blassen eingefallenen Wangen** in seinem Bett zwischen aufgesteckten Kissen.	There he sat with **pale, sunken cheeks** in his bed between propped-up pillows.
Basic attitude	Wir wissen, es **lag nicht in seinem Wesen, zu rennen, unanständige Eile war ihm fremd**, seine Korpulenz verbot ihm geradezu, Sprünge zu machen.	We know that it was **not in his nature to run, indecent haste was foreign to him**, his corpulence positively forbade him to jump.
Charisma	Ist sie nicht wie die Morgenröte **lieblich**?	Is she not **lovely** like the dawn?
Nationality/place of origin	Denn sie **war aus Sachsen**.	For she **was from Saxony**.
Standard of living	Als er wieder zurückkam, kündigte er das Atelier, unsere **hübsche, große Wohnung**, und mietete eine viel kleinere.	When he came back, he gave up the studio, our **pretty, large apartment** and rented a much smaller one.
Toe/foot/leg	Und eines Tages raffte sie ihr Kleid bis fast zum Knie: »Habe ich nicht ein **schönes Bein**, Albertchen?«	And one day she gathered her dress up almost to her knees: “Don’t I have **beautiful legs**, Albertchen?”
Religion/politics	sie waren Freigeister, ohne sich so zu nennen oder es auch nur zu wissen, der **Vater Lutheraner**, die **Mutter Katholikin**.	They were free spirits, without calling themselves that or even knowing it, the **father was Lutheran**, the **mother was Catholic**.
Piece of clothing	Herr Ewald Brock knöpfte seinen **Frack** auf, strich sich über den Leib und sagte: »Mehlsuppe!«	Mr. Ewald Brock unbuttoned his **tailcoat**, stroked his chest and said, “Flour soup!”
Accessoires	Auch schlug er mit seinem **Spazierstock** eine steile Terz in die Luft.	He also struck a steep third in the air with his **walking stick**.
Trunk/shoulder	Der **Rücken schien hart** zu sein;	His **back seemed to be hard**;

Table 10: Examples for each fine-grained category, in the original German and an English translation.

C Inter-Annotator Agreement

Table 11 shows the full list of agreement values (Fleiss κ), overall, for the coarse-grained and for the fine-grained categories.

Category	Fleiss κ	z-score
Overall coarse-grained	0.935	21.99
Overall fine-grained	0.232	24.135
Age	0.856	11.252
Character trait	0.94	12.368
Clothing	1.00	13.153
Physiognomy	0.961	12.636
Role	0.93	12.234
Accessories	0.388	10.664
Charisma	0.272	7.485
Occupation	0.418	11.487
Relationship	0.452	12.431
Sex	0.375	10.316
Family	0.398	10.947
Mind/habitus	0.392	10.782
Face	0.352	9.682
Basic attitude	0.238	6.541
Piece of clothing	0.401	11.029
Head/hair	0.396	10.888
Body/health	-0.001	-0.036
Height/stature/weight	0.438	12.048
Standard of living	0.438	12.036
Numerical age	0.354	9.74
Role with connection to age	0.386	10.612
Scalar age	0.412	11.315
Social status	0.497	13.675
Type	-0.001	-0.036

Table 11: Fleiss κ per fine-grained category and z-score. All values are statistically significant ($p \approx 0$) except for “body/health” and “type” ($p = 0.971$).