Annotating Spatial Descriptions in Literary and Non-Literary Text

Emilie Sitter, Omar Momen, Florian Steig,
Berenike Herrmann, and **Sina Zarrieß**
CRC 1646 – Linguistic Creativity in Communication
Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{emilie.sitter,omar.hassan,f.steig,berenike.herrmann,sina.zarriess}@uni-bielefeld.de

Abstract

Descriptions are a central component of literary texts, yet their systematic identification remains a challenge. This work suggests an approach to identifying sentences describing spatial conditions in literary text. It was developed iteratively on German literary text and extended to non-literary text to evaluate its applicability across textual domains. To assess the robustness of the method, we involved both humans and a selection of state-of-the-art Large Language Models (LLMs) in annotating a collection of sentences regarding their descriptiveness and spatiality. We compare the annotations across human annotators and between humans and LLMs. The main contributions of this paper are: (1) a set of annotation guidelines for identifying spatial descriptions in literary texts, (2) a curated dataset of almost 4,700 annotated sentences of which around 500 are spatial descriptions, produced through in-depth discussion and consensus among annotators, and (3) a pilot study of automating the task of spatial description annotation of German texts. We publish the codes and all human and LLM annotations for the public to be used for research purposes only.[1]

1 Introduction

Literary and non-literary texts are full of descriptions that help readers see, hear, feel, smell, and even taste what is happening in a story or text, making the places and entities experiencable. While the analysis of literary text has become an important area of annotation studies, existing work typically targets narrative elements, such as characters or plot structure (Bethard et al., 2012; Reiter, 2015; Bamman et al., 2020; Zehe et al., 2021; Jahan et al., 2021; Reiter et al., 2022; Soni et al., 2023). In the domain of non-literary text, a lot of recent NLP work deals with multimodal image descriptions scraped from alt-texts on the web or collected via human annotations, cf. (Young et al., 2014; Sharma et al., 2018; Pont-Tuset et al., 2020; Garg et al., 2024; Alaçam et al., 2024). However, to our knowledge, no tool or dataset distinguishes between descriptive and non-descriptive language and identifies descriptions in naturally occurring text. In this work, we present an approach to annotating and detecting descriptions in unimodal, literary, and non-literary text. To give our study a concrete target and domain, we focus on descriptions of space.

Since the 1990s, the concept of space has gained increasing attention in the cultural and social studies (Döring and Thielmann, 2008). In linguistics and NLP, the analysis of spatial language in text has received moderate but continuous attention. To date, existing work on annotations of spatial language mainly aimed at detecting mentions of spatial entities (named entity recognition) or other spatial concepts, like paths or trajectories (Pustejovsky et al., 2015; Pustejovsky, 2017).

This work focuses on identifying sentences describing static space. The following sentence is an example of a spatial description in a story that works without naming any named spatial entities:

(1) Auf dem zertretenen Rasen zwischen Haus und Zaun, roh gezimmert, stand ein länglicher Tisch mit Bank und Sesseln.[2]
On the trampled lawn between the house and the fence, rough-hewn, was an oblong table with a bench and chairs.

In literary texts in particular, such descriptions are a fundamental unit for creating a space of action and opening up a world to the reader by routing the narrative in a physical environment. Despite the increasing interest in space and spatial descriptions,

[1] https://github.com/emilie-si/LAW2025-Descriptions

[2] Arthur Schnitzler: Doktor Gräsler, Badearzt (1917)

identifying them in a natural context—in our study, novels or travel reports—remains a challenge. The paper contributes to the broader goal of understanding spatial and descriptive language in various textual domains and improving its automatic detection. We propose a set of annotation guidelines to extract spatially descriptive sentences from literary and non-literary texts beyond self-evident cases. As examples we use the two German corpora KOLIMO (Herrmann and Lauer, 2018; Horstmann, 2019) and Wikivoyage (Nolda, 2024; Wikimedia Foundation Inc., 2025).

Based on samples extracted from these two corpora, we created a set of annotated sentences. To ensure that all annotators' perspectives are considered, we systematically discussed the cases of disagreement. A final label was assigned based on the mutual agreement of all annotators on a plausible classification. Since human annotations are expensive and time-consuming, we also explore how to automate this annotation task. Based on the manually annotated dataset, we test the ability of LLMs to identify spatial descriptions. In doing so, we aim to contribute to a more comprehensive understanding of spatial language processing.

2 Background: Descriptions and Space

2.1 Descriptions

We draw on background from different disciplines to develop our approach to annotating descriptions. Since our main focus is on literary text, we rely on work from literary studies (Ronen, 1997; Hahn et al., 2025), digital humanities (Herrmann et al., 2022; Schumacher, 2023), and psychology (Draschkow and Võ, 2017; Henderson and Hollingworth, 1999).

It can be assumed that humans generally have an intuitive understanding of what is descriptive (Wolf, 2007; Nünning, 2007). Depending on the domain and genre of a text, spatial conditions can be presented in different contexts and for different reasons. The primary function of spatial descriptions is to convey spatial information (Ryan, 2012). They enable readers to build a mental figuration of spatial information (Denis, 2008, 2018) and serve as a building block for constructing narrative space (Dennerlein, 2009; Wolf, 2007).

The boundary between narrative and descriptive is more than often fluid. We are thus taking up the long-standing question of how to reliably distinguish between narrative and descriptive (Mosher, 1991; Ronen, 1997; Wolf, 2007). According to Wolf, a distinction can be made by "the presence or absence of the core elements of typical narratives: motivated actions that involve anthropomorphic agents, are interrelated not only by chronology but also by causality and teleology and lead to, or are consequences of, conscious acts or decisions, frequently as results of conflicts" (Wolf, 2007). Similarly, for Dennerlein "uneventfulness and the communication of stable properties of a spatial situation" are the central criteria of spatial descriptions (Dennerlein, 2009, own translation).

However, there are countless cases in which these two criteria are either not exclusively or not fully met (Ronen, 1997). This work shows how we deal with such cases.

2.2 Spatial Frames

The sentences relevant in our annotation task should describe visually cohesive spaces with scenic quality. In the literary studies, Ruth Ronen's concept of "spatial frames" refers to this relatively restricted sub-area of space: spatial frames are "the actual or potential surroundings of fictional characters, objects and places" (Ronen, 1986). Spatial frames encompass only the (potential) environment of a narrator or the characters in a story: everything that could be perceived as being "here" during narration and where an action can (potentially) take place (Zoran, 1984; Ryan et al., 2016). The notion of spatial frames as "shifting scenes of action" Ryan et al. (2016) highlights the scenic nature of spatial frames.

The entire space in which a story takes place can be understood as a series of many individual spatial frames (Zoran, 1984). Spatial frames are different to specific locations. They represent particular, immovable points in space that can be localized either on a real map or on the map of a story world (Schumacher, 2023; Ryan et al., 2016). Places become spatial frames as soon as they convey more meaning than a mere geographical location on a map.

Grounding our description identification approach on Ronen's (1986) concept of Spatial Frames has certain advantages. It excludes instances of spatial language that do not exactly describe spatial conditions, such as route descriptions or mere geographical and factual information (as in "*Berlin is the capital of Germany*"). But, compared to more restrictive concepts, it includes any kind of space as long as action could take place

there within a story ("*Berlin is big and noisy.*"). Spatial Frames in a story do not only encompass a character's actual spatial surroundings but everything that, within the story, can *potentially* be their environment (Ronen, 1986). Since we annotated isolated sentences without context, it cannot always be judged what would be an actual surrounding in a story and what is, for instance, only imagined, dreamed, or described from afar. Spatial frames comprise exactly the section of spatial language that we want to capture in our annotation task.

2.3 Scenes

Objects share some qualities with spatial frames, such as their three-dimensionality and perceptibility (they can be experienced on various levels, such as visually, acoustically, haptically). However, in contrast to scenes in which we can be embedded and events can take place, we can look at discrete objects only from an outside point of view (Henderson and Ferreira, 2004).

Drawing an analogy between textually described scenes and visually depicted scenes (in real life or in photographs), we rely on the concept of Scene Grammar (Draschkow and Võ, 2017; Võ and Wolfe, 2013; Võ et al., 2019; Wolfe et al., 2011) to distinguish objects from scenes. Assuming that scene perception functions in a similar way to language perception, it serves as an approach for understanding the generation of mental models of described scenes. Scene Grammar comprises the environmental rules that help us to recognize real-world visual scenes at first glance by only coarse spatial information (Draschkow and Võ, 2017; Võ et al., 2019; Oliva, 2005).

According to Scene Grammar, a combination of individual, static anchor objects (e.g., shower, washbasin, toilet) and smaller-scale local objects attached to anchors (e.g., towel, soap bar, toilet paper) forms a complete scene (e.g., bathroom) (Võ et al., 2019; Draschkow and Võ, 2017; Oliva, 2005). In our annotation task, we rely on Scene Grammar to exclude descriptions of anchor objects on their own (such as "*The towel is red.*"). However, a combination of explicitly ("*Next to the clean shower, there is a red towel.*") or implicitly ("*The bathroom is clean.*") described individual objects indicates that the subject of the description is a scene. We can then consider it a spatial frame.

3 Annotation Procedure

This section introduces the set-up of our annotation task, the procedures for guideline development and data curation, as the final annotation guidelines.

3.1 Approach

We asked our annotators to identify spatial descriptions on the level of complete, isolated sentences (we do not consider passages describing space that are shorter or longer than exactly one complete sentence). The annotators' task was to make a binary distinction, i.e., whether an instance is a spatial description or not. Moreover, annotators could annotate instances as "unclear" and could add a comment explaining their uncertainty. All sentences were annotated independently by one of the paper's authors and two out of a group of four in-lab trained annotators.

3.2 Iterative Guideline Development

We followed Reiter's (2020) proposed methodology for developing annotation guidelines. This approach aims to develop generic but precise guidelines for the practical annotation of a phenomenon that has already been described theoretically.

We started the guideline development for the literary data, assuming that it is more difficult to identify static spatial descriptions in literary and narrative than in non-narrative texts. The initial round of annotations was conducted in a relatively open manner, aiming to better understand the phenomenon and to identify ambiguities and challenges. The guidelines were then iteratively developed and refined based on existing research on the subjects of space, description, and scenes. They are formulated in bullet points and contain examples for all cases described (Reiter, 2020; Reiter et al., 2019).

After annotating a subset of sentences, we discussed the individual diverging samples and further sharpened the guidelines as reported in Section 3.4. If annotators chose different categories or the label "unclear" due to a lack of clarity in the guidelines, these were adjusted accordingly. All annotators were informed of the update.

3.3 Data Curation

To obtain a curated ground-truth dataset, we took into account all annotators' subjective decisions and re-evaluated divergent annotations through discussion. A final label was assigned based on mutual agreement. The aim was to finally select categories

as comprehensible and acceptable to as many annotators as possible. Guideline adjustments of later annotation iterations were incorporated retroactively into previously annotated subsets. This procedure ensured the creation of a curated dataset with the most appropriate categories.

Please refer to Section 5 for further analysis of annotator agreement and Section 7 for further discussion.

3.4 Annotation Guidelines

This section summarizes the guidelines that were iteratively developed for identifying spatial descriptions in literary text.

1. Spatial descriptions describe "spatial frames": any space that can potentially be a character's immediate environment in a story (Ronen, 1986). They describe an actually perceptible scene (2-a) instead of, for instance, only background knowledge about a location (2-b).

(2) a. There was a scent of flowers in the pretty looking garden. (✓)
b. The garden was redesigned last year. (✗)

2. Spatial descriptions must contain information about the spatial and perceptible environment at a certain place. Spatial frames can be captured by describing what can be perceived at a certain point in space. Rather than just mentioning a spatial frame (3-b), there has to be some descriptive element (3-a).

(3) a. This forest is dark. (✓)
b. This is a forest. (✗)

3. Spatial descriptions can also convey acoustic, tactile, olfactory, or other sensory signals that contribute to the perception of space (4-a) (Wolf, 2007). Describing the spatial frame not necessarily requires visual sensations, as we can infer the spatial conditions through these other sensory modalities (Dennerlein, 2009).

(4) a. In the basement it was cold and a mildewy scent hung in the air. (✓)

4. Spatial descriptions describe a scene (5-a) instead of a single object (5-b). We can define a scene as an arrangement of two or more implicitly or explicitly mentioned independent elements in a semantic relationship.

(5) a. There is a green bottle on the table. (✓)
b. My bottle is green. (✗)

5. An isolated sentence must not contain any unresolved references to previous text (e.g. pronouns) (6-b). Any spatial description can be understood without any further textual context (6-a).

(6) a. The living room was furnished tastefully. (✓)
b. It was furnished tastefully. (✗)

6. Descriptions do not report any action. The described space is static, its properties are stable over time. There is no unique, temporary action (which would often be expressed by a verb for a spontaneous, individual action or movement, such as "walk") at the time of description of the space (7-c). Descriptive parts of sentences that are embedded in narrative sentences Schumacher (2023) are not relevant for our annotation task. The following exceptions can be made: a) typical and recurring actions of generic actors who are not individual characters in the passage (Dennerlein, 2009) (7-a) and b) the act of perception reported while describing space (by verbs of perception, such as "see" or "hear") (7-b).

(7) a. Shibuya Crossing is constantly filled with pedestrians. (✓)
b. We saw the small bridge that crosses the river. (✓)
c. We crossed the river over a small bridge. (✗)

7. For the description of generic, natural phenomena and light, we apply a WIDLII (*When In Doubt, Leave it In*) approach (Steen et al., 2010). With natural phenomena (weather and wind, tides and waves, daylight phases, sunrises and sunsets, clouds, light from lamps or candles) there is usually some kind of movement: waves roll over the water, clouds drift across the sky, the sun rises or sets. The described natural phenomena must not contain a narrative and have to be generic and repetitive instead of one-off movements (8-a).

(8) a. The sun sank, painting the horizon a breathtaking red. (✓)

8. Only concrete space is of interest to us. Described space can be real or fictional, imaginary, remembered, phantastic, or dreamed, as long as it is not purely metaphorical or an abstraction of a character's mental processes (9-a).

(9) a. There was a maze of thoughts tangled up in my mind. (✗)

9. The spatial descriptions must be complete German sentences, but a verb is not necessarily required (10-a).

(10) a. Colorful flowers, ripe fruit, large trees in the garden. (✓)

4 Spatial Descriptions Dataset

Our annotation work resulted in a dataset of spatial descriptions extracted from two fundamentally different German corpora of literary and non-literary texts: KOLIMO and Wikivoyage. KOLIMO, the "Corpus of Literary Modernism", has its focus on 19th century fiction (Herrmann and Lauer, 2018; Horstmann, 2019). The copyright on these texts has expired, and they are public domain. KOLIMO is a convenient literary corpus because of its size and its availability in digital form with extensive metadata. As a non-literary counterpart, we chose Wikivoyage, an online travel guide, as we expected to find many spatial descriptions there (Nolda, 2024; Wikimedia Foundation Inc., 2025). The German version of Wikivoyage is distributed under the CC BY-SA 4.0 license.

We developed our guidelines for spatial descriptions primarily based on KOLIMO. As a non-literary counterpart that is highly different not only in genre but also in its time of origin, Wikivoyage enables us to explore the extent to which the annotation scheme can be transferred to another domain.

For annotating on the sentence level, the full texts required some preprocessing. We excluded texts shorter than 10 sentences, assuming that it is unlikely that authors will dedicate complete sentences to exclusively describe spatial surroundings in very short texts. We eliminated incomplete sentences and only included sentences that begin with a capital letter and end with a punctuation mark.

	KOLIMO	**Wikivoyage**
Time Span	1850–1939	2012–2024
# Texts	43,012	20,195
# Filtered Texts	14,901	17,781
# Filtered Sentences	7,783,056	876,775
# Annotated Sentences	3854	800
Spatial Descriptions Ratio	8.4%	20%

Table 1: Statistics of the two corpora used in our study.

Bullet points, as they can be found in Wikivoyage, inherently indicate the beginning of a sentence and, therefore, cannot appear within a sentence. Moreover, only sentences with a minimum length of five words are considered for annotation. Table 1 reports the size of the complete dataset.

For better comparability between the two subsets, we pre-filtered the data. For each corpus, we determined the 10 most frequent non-named spatial entities (by lemma) (Kababgi et al., 2024) based on a list of spatial entities generated by Herrmann et al. (2022). Inflected forms or spatial entities as part of compound words (as they are frequent in German) were taken into account as far as possible (see Appendix A). We condensed the datasets to only sentences that contain one or more of the 10 most frequent spatial entities.

Pre-filtering definitely contributed to the proportion of spatial description among all annotated sentences, as reported in Table 1. We ensure that all sentences contain at least one spatial entity and, therefore, are spatial to some degree. Otherwise, at least in the literary data, a lower proportion of descriptions would be expected (Ronen, 1997).

5 Analysis: Agreement and Challenges

5.1 Quantitative Evaluation

For a quantitative evaluation of annotator agreement, three annotators independently annotated subsets of 300 sentences in random order. Disagreement cases were discussed individually and used to further refine the annotation guidelines and to train the annotators (see Section 3.4). Starting with literary sentences, we measured their Inter-Annotator Agreement (IAA) by Krippendorff's alpha (Krippendorff, 2013) and the F1 score in every iteration, as shown in Table 2. Instances annotated as "unclear" were counted as "not a spatial description" since our focus is on clear cases of descriptions. The highest achieved Krippendorff's Alpha in the best annotation iteration (iteration 2) is .66. Table 2 also shows that the continuous adaptation of

	It. 1 (Lit.)	It. 2 (Lit.)	It. 3 (Lit.)	It. 4 (Non-lit.)
# Sent.	294	295	300	300
A1-A2-A3 (K–α)	.63	.66	.60	.44
A1-A2 (F1)	.70	.65	.65	.58
A1-A3 (F1)	.67	.69	.74	.58
A2-A3 (F1)	.61	.72	.56	.40
A1-LLM (F1)	.64	.62	.71	.13
A2-LLM (F1)	.62	.73	.53	.12
A3-LLM (F1)	.51	.64	.67	.09
Curated-LLM (F1)	.70	.65	.70	.08

Table 2: Agreement between annotators and best LLM (Qwen2.5:32B with long English prompt (EN-long)). The table reports the agreement between the annotators and the annotators and the model in four iterations (It. 1 to It. 4) of annotating 300 sentences across both Literary and Non-literary datasets. (Some sentences of these sets were used to develop the prompt and are therefore not considered in this evaluation.)

the guidelines and excessive training of the annotators resulted in the agreement decreasing again in iteration 3.

The guidelines for literary text were slightly adapted to account for the non-literary corpus. These sentences exhibit a different structural composition. Surprisingly, they were not as easy to identify with the existing set of rules, which is again reflected in the decreasing IAA of iteration 4. For the pilot study, we tested the applicability of the existing rules to the non-literary texts, but these need to be further adapted in order to consistently identify spatial descriptions in this corpus.

5.2 Qualitative Evaluation: Literary Text

Literary text often allows for more than one correct interpretation (Gius et al., 2019; Gius and Jacke, 2017; Amidei et al., 2018). A particular challenge in our corpus is to distinguish the narrative or partially narrative sentences from those that are exclusively descriptive. Often, some degree of subjectivity underlies the annotation, as in the following examples:

In Example 1 in Appendix B, the annotators disagreed concerning the concreteness of the described space. One annotator was arguing that in this case the city is a concrete space that is actually described, while others assumed that the sentence reflects the mental state of the narrator.

As for Example 2 in Appendix B, the annotators could not agree whether the sentence can be considered as an action, or if sleepers lying on the earth should correctly be interpreted as a stable property of the described space.

Annotators also interpreted Example 3 in Appendix B differently. It was not clear whether describing what the room *not* is would be sufficient or too little information for a spatial description.

5.3 Qualitative Evaluation: Non-literary Text

In Wikivoyage, sentences with specific and temporary actions are rare, but the corpus contains many geographical descriptions, route descriptions, and street courses. These are spatial in a certain way but do not exactly represent spatial frames. Descriptions of mere geographical locations only provide information on where a specific place (a named entity) can be located on a map, as in Example 4 in Appendix B. If only slightly more spatial information is provided (as in Example 5) it becomes unclear whether the passage should still be classified as a geographical description or already constitutes a spatial frame.

Route descriptions describe the way from one to another location and possible landmarks along the way (Denis, 2018). These kinds of descriptions do not correspond to the immediate, perceptible surroundings at a specific location and can therefore be excluded from our annotation scheme (see Example 6 in Appendix B). However, when they also describe spatial properties, as in Example 7, they could be interpreted as spatial frames.

In the literary corpus, the vast majority of sentences is complete. Ellipses can be considered complete sentences. In literary text, they can serve as rhetorical devices (see Example 8 in Appendix B). In Wikivoyage, on the other hand, we found sentences without any verbs, serving as enumerations, abbreviations, or points on a bullet list (as in Example 9 in Appendix B). By definition, these are complete sentences as they begin with a capital letter and end with a punctuation mark. As long as there is a semantic relationship between the listed elements, the absence of a verb does not necessarily make a sentence an uninterpretable array of random objects (Henderson and Ferreira, 2004). To prevent doubts as to whether it is even possible to describe without a verb, the guidelines had to be adapted to state explicitly that the occurrence of a verb is not a decisive criterion for annotation.

6 Pilot Study: Automatic Annotation

Our aim is to eventually have a larger dataset of spatial descriptions across different textual domains. To this end, we carried out a prompting experiment with LLMs to classify the literary and non-literary sentences in our dataset (§ 4) in a zero-shot setting.

6.1 Experimental Setup

To track the effect of the variables in this experiment (input prompt, model family, and model size), we used four different prompts and seven different models to classify the 3854 literary and 800 non-literary sentences, resulting in 28 automatic annotations for each sentence. We measured the performance of these annotations using the human annotations as the ground truth.

We developed four different prompts in English and German, with varying levels of detail based on the annotation guidelines. We chose to use the German prompt only in the *long* version, as there were no significant differences between languages in the other levels of detail. Then we explored the prompts' performance on 70 randomly selected sentences from the set of annotated literary sentences. These 70 sentences were not considered in the further evaluation. The prompts were modified slightly for the non-literary sentences (see Appendix C).

LLMs have been evolving rapidly, and no single model offers the best performance across the board. Different model families and sizes each have their advantages and disadvantages. To account for this, we tested several different models: GPT-4o, one of OpenAI's current proprietary LLMs; Gemma2 and Qwen2.5, two open-source LLMs. For each of these two open-source models, we tested 3 different model sizes, ranging from 2B to 32B parameters. We report the experiment's settings in Appendix D.

We could successfully get a clear answer as (YES/NO) for almost all the responses in our prompting experiments; only in very few cases we had to manually look at the response to figure out the answer. Eventually, we transformed all the responses into binary labels. This enabled us to evaluate the performance of the 28 model-prompt variants against the human annotations. We measured accuracy, precision, recall, and F1 score of each variant. Additionally, we report the ratio of sentences predicted as spatial descriptions to the total number of sentences in the dataset for each variant, considering that the ratio in human annotations (prior probability) is .08 for literary texts and .20 for non-literary texts.

6.2 Results

We report the results of the top five models (according to F1 score on literary sentences) in Table 3. The results of all model-prompt variants for the literary and non-literary dataset are reported in Appendix E. Results of the literary dataset in Table 3 show that all models achieve high accuracies (.82-.95), but face a severe precision-recall tradeoff, resulting in lower F1 scores (.45-.67). Most models show a low ratio of predicting descriptions, roughly aligning with the low ratio of descriptions in the human annotations. We notice that the best-performing models on the literary dataset show very different results on the non-literary dataset. The accuracies deteriorate by 10-15 points, and the models are either extremely restrictive in classifying sentences as descriptions or make a lot of mistakes when being less restrictive (row 3).

The variants with the highest F1 for literary sentences (.67, .64, .57) are (Qwen2.5:32B, EN-long), (GPT-4o, EN-long), and (Qwen2.5:7B, EN-medium) respectively. (Qwen2.5:32B, EN-long) is better at precision, while precision and recall of (GPT-4o, EN-long) are more balanced. As for model families, Qwen is performing generally better than Gemma, and it also outperforms the closed-source representative GPT-4o. Larger size does not always guarantee (significantly) better performance across each model family, as highlighted by Qwen2.5:7B results, which are relatively better than those of the 32B variant at the (EN-medium) prompt variant. However, we notice that the 3B versions of Qwen2.5 chose NO for all sentences, resulting in zero true positives, and hence zero precision, recall, and F1. For prompt variants, generally, the longer detailed prompts perform better than the shorter ones, and the German prompt does not improve over the English version. Exceptions show that the 7B version of Qwen performs better with briefer prompts than detailed ones, and that Gemma models perform better with the German prompt than the English one.

In Table 2, we compare the F1 scores between annotator pairs and between each annotator and our best-performing model-prompt variant on the literary dataset (Qwen2.5:32B, EN-long). The results show that the F1 score of the automatic annotations falls in the same range as the F1 scores of the annotator pairs. In the literary dataset, the val-

		Literary Dataset					Non-Literary Dataset				
Model	**Prompt**	**Acc.**	**P**	**R**	**F1**	**Rat.**	**Acc.**	**P**	**R**	**F1**	**Rat.**
Qwen2.5:32B	EN-long	**.95**	**.83**	.56	**.67**	.06	.81	1.0	.06	.12	.01
GPT-4o	EN-long	.94	.64	.63	.64	.08	.84	.97	.19	.32	.04
Qwen2.5:7B	EN-med.	.93	.56	.57	.57	.09	.76	.40	.42	.41	.21
Gemma2:27B	DE-long	.86	.37	.86	.52	.20	.84	.81	.26	.40	.06
Gemma2:9B	DE-long	.82	.31	**.88**	.45	.24	.84	.87	.21	.34	.05

Table 3: Evaluation results of the top five models according to F1 on the literary dataset. We selected only the best-performing prompt variant for each of these models. We report **Acc**uracy, **P**recision, **R**ecall, **F1**, and **Rat**io of predicted sentences as spatial descriptions to the total number of sentences in each dataset (literary dataset: 3784 sentences; non-literary dataset: 800 sentences).

ues range between .56 and .74 for annotator pairs, and between .51 and .73 for LLM-Human pairs. For non-literary texts, the values are lower for both annotator pairs and LLM–human pairs, with extremely low F1 scores for the latter. These low scores on the non-literary dataset suggest a significant change in task difficulty for LLMs across different genres. They highlight the need for genre-specific prompts, reflecting the varying annotation guidelines between genres.

In summary, the pilot study illustrates the usability of LLMs at the task of classifying sentences as spatial descriptions. For the literary sentences, they produce annotations with an acceptable degree of accuracy and a precision-recall trade-off, considering the inherently uncertain nature of the task. We found that the (Qwen2.5:32B, EN-long) model-prompt variant yields predictions that agree the most with human annotations for literary texts. Moreover, we found that no single model-prompt variant could perform consistently well across both literary and non-literary datasets. The guidelines and then the prompts were developed for the literary sentence. The transfer to Wikivoyage—an experiment as part of the pilot study—demonstrated that the guidelines and prompts have to be adapted to obtain reliable annotations, taking into account the different textual domains and times of origin.

It is also important to note that the pilot study was conducted on the subset of data restricted to sentences describing specific spatial entities reported in § 4. Therefore, the extent to which our prompts generalize to the full corpora remains uncertain at this stage.

7 Discussion

Natural language and especially literary text is inherently complex and often ambiguous. In our aim to identify spatial descriptions, we encountered several sources of disagreement. Apart from uncertainties in the texts themselves, disagreement also resulted from unclear cases within the annotation guidelines and practical factors such as annotator error. In this section, we discuss the major reasons for annotator disagreement. Unresolvable ambiguities within the data itself are the most prominent factor for disagreement. Isolated sentences do not always provide clear evidence as to whether they constitute a spatial description according to our definition. (See, for instance, Example 10 in Appendix B: without context, our annotators could interpret it as a description of an actual, spatial scene as well as a pure abstraction and therefore not spatial. Examples 11 and 12 were ambiguous for our annotators due to the polysemy of certain words.) Pavlick and Kwiatkowski's (2019) results, on the other hand, suggest that an increased amount of context would not necessarily contribute to an increased IAA. We therefore assume that there will always be at least a certain level of disagreement between annotators simply due to the polyvalence of literary text (Gius and Jacke, 2017).

When the guidelines lack precision, however, it can result in fuzziness and different interpretations not of the text itself, but of the annotation scheme. Gius and Jacke (2017) claim that any fuzziness in the categorization must be minimized as much as possible. The inherent polyvalence of the texts does not justify ambiguity in the category definitions. On the other hand, it is generally not possible to formulate guidelines that unambiguously account for 100% of all cases (Reiter et al., 2019). Our

attempts to make the guidelines as precise as possible resulted in a detailed seven-page document. Amidei et al. (2018) warn of guidelines becoming too narrow and restrictive. They would be at risk of failing to capture the variability and polyvalence inherent to human language. In iteration 3 of our annotation, we had the most extensive list of guidelines in use. As Table 2 reports, the agreement between the annotators decreased. The guidelines would have covered most of the cases, but the cognitive load for the annotators was too high and they were too narrow to generalize well across our data.

A third and minor, but still a noticeable reason for an imperfect IAA was human errors (Pavlick and Kwiatkowski, 2019). When processing a large number of individual sentences in succession, the cognitive effort of the annotators was considerable and could occasionally lead to the selection of incorrect categories.

We argue that certain levels of disagreement are not only unavoidable but even indicative of the nuanced nature of descriptive and spatial language. We did not expect perfect agreement between the human annotators and even less between humans and LLMs. Instead, the objective was to produce a curated dataset of spatial descriptions in which any ambiguity arises solely from legitimate differences in the interpretation of language, accounting for the subjectivity of the individual annotators (Reiter et al., 2019; Amidei et al., 2018). The annotation process provided valuable insights into how humans interpret descriptive and spatial language and how annotation guidelines mediate this interpretation.

In general, we observe that the task of description annotation features a certain amount of subjectivity, resulting in label variation in our data. While traditional NLP paradigms aimed at eliminating human label variation as much as possible, recent work argues for embracing rather than excluding or ignoring it (Plank, 2022; van der Meer et al., 2024). By making the different iterations of our annotations and guidelines available, we also hope to contribute to this emerging line of research.

Conclusion

This work presents an approach to identifying spatial descriptions in literary text. A group of human annotators and of LLMs annotated individual sentences to determine whether they are spatial descriptions. While space and spatiality are topics that have received considerable attention in the (digital) humanities, literary studies, and, to some extent, in computational linguistics, this work is among the first to explicitly focus on the systematic identification of descriptions. We propose a set of annotation guidelines for spatial descriptions and report the performance of multiple LLMs in this annotation task. Our analyses revealed several systematic challenges for the manual and automatic annotation of descriptions, such as annotator subjectivity in assessing semantic aspects like concreteness and ambiguities as well as issues with substantial differences between datasets and class imbalance. A valuable next step could now be to investigate the impact of additional in-context examples or task-specific fine-tuning. Moreover, the relatively low agreement score of .44 for non-literary texts indicates that the annotation guidelines require further adjustment for this domain.

Limitations

One major limitation of this work is extending the existing annotation scheme to non-literary text. There are substantial differences between the two corpora we worked with not only in their textual structure but also in the time period they cover. The guidelines developed for literary text were less applicable to non-literary texts than expected. It turned out that for a reliable annotation of non-literary sentences, new guidelines and completely new prompts, along with a re-training of the annotators, would have been required.

Moreover, KOLIMO covers the literary domain (German-language texts from the late 19th century and early 20th century) much more extensively than Wikivoyage represents the non-literary domain. We are aware that travel reports cannot be equated with a general "non-literary" language, which includes many more text types and genres.

A possible extension of the dataset for a follow-up study could therefore include other corpora, especially from the non-literary side, in order to investigate annotators' and LLM's abilities to identify spatial descriptions in this data. However, also corpora of other languages than German could be of interest.

Our approach to counting the most frequent spatial entities is inherently flawed, as Herrmann et al.'s (2022) spatial entity list is by far not comprehensive. It was generated to cover literary fiction from the 19th and 20th century and therefore works

better for KOLIMO than the contemporary texts in Wikivoyage. For instance, "Flughafen" ('airport') is not part of the list, however, due to our matching of compounds, this entity will be considered as an instance of "Hafen" ('harbor', 'port'). Moreover, it comprises only single words, while spatial entities could also be expressed as nominal phrases (see e.g., Barth (2021)).

A better approach instead of the list and regular expressions would be to use a neural model for a proper counting of the most frequent entities and then selecting the relevant sentences. However, at the time of creating the data set, we were not aware of any model for German that could automatically extract all relevant spatial entities from our large datasets. Moreover, for the time being we only aimed to control the dataset for our annotators in order to avoid annotating sentences entirely at random. The purpose of the pre-filtering is not to identify spatial sentences but to create a set of filtered candidate sentences that is more meaningful than a set composed of completely random corpus sentences.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A05.

References

Özge Alaçam, Ronja Utescher, Hannes Grönner, Judith Sieker, and Sina Zarrieß. 2024. WikiScenes with descriptions: Aligning paragraphs and sentences with images in Wikipedia articles. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 93–105, Mexico City, Mexico. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Florian Barth. 2021. Konzept und klassifikation literarischer raumentitäten. pages 1281–1293. ISBN: 9783885797012 Publisher: Gesellschaft für Informatik, Bonn.

Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating story timelines as temporal dependency structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2721–2726, Istanbul, Turkey. European Language Resources Association (ELRA).

Michel Denis. 2008. Assessing the symbolic distance effect in mental images constructed from verbal descriptions: A study of individual differences in the mental comparison of distances. 127(1):197–210.

Michel Denis. 2018. *Space and spatial cognition: a multidisciplinary perspective*. Routledge.

Katrin Dennerlein. 2009. *Narratologie des Raumes*. De Gruyter. Publication Title: Narratologie des Raumes.

Dejan Draschkow and Melissa Le-Hoa Võ. 2017. Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. 7(1):16471.

Jörg Döring and Tristan Thielmann. 2008. *Einleitung: Was lesen wir im Raume? Der Spatial Turn und das geheime Wissen der Geographen*, pages 7–46. transcript Verlag, Bielefeld.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127, Miami, Florida, USA. Association for Computational Linguistics.

Evelyn Gius and Janina Jacke. 2017. The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. 11(2):233–254. Publisher: Edinburgh University Press.

Evelyn Gius, Nils Reiter, and Marcus Willand. 2019. A shared task for the digital humanities chapter 2: Evaluating annotation guidelines. 4(3).

Kurt Hahn, Anne-Kathrin Reulecke, Steffen Schneider, and Julia Zimmermann, editors. 2025. *Descriptio*, 1 edition, volume 263 of *Litterae*. Rombach Wissenschaft, Baden-Baden.

John M. Henderson and Fernanda Ferreira. 2004. Scene perception for psycholinguists. In *The Interface of Language, Vision, and Action*. Psychology Press. Num Pages: 58.

John M. Henderson and Andrew Hollingworth. 1999. High-level scene perception. 50:243–71.

J. Berenike Herrmann, Joanna Byszuk, and Giulia Grisot. 2022. Using word embeddings for validation and enhancement of spatial entity lists.

J. Berenike Herrmann and Gerhard Lauer. 2018. Korpusliteraturwissenschaft. zur konzeption und praxis am beispiel eines korpus zur literarischen moderne. 92:127–156.

Jan Horstmann. 2019. KOLIMO: Korpus der literarischen moderne.

Labiba Jahan, Rahul Mittal, and Mark Finlayson. 2021. Inducing stereotypical character roles from plot structure. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 492–497, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Kababgi, Giulia Grisot, Federico Pennino, and J. Berenike Herrmann. 2024. Recognising non-named spatial entities in literary texts: a novel spatial entities classifier. In *CHR 2024: Computational Humanities Research Conference*, pages 472–481.

Klaus Krippendorff. 2013. Computing krippendorff's alpha-reliability.

Harold F. Mosher. 1991. Toward a poetics of "descriptized" narration. 12(3):425–445. Publisher: Duke University Press, Porter Institute for Poetics and Semiotics.

Andreas Nolda. 2024. Wikivoyage-korpus: Korpusquellen der deutschen sprachversion von wikivoyage im TEI-format.

Ansgar Nünning. 2007. Towards a typology, poetics and history of description in fiction. In Walter Bernhart and Werner Wolf, editors, *Description in Literature and Other Media*, volume 2 of *Studies in Intermediality (SIM)*, pages 91–128. Brill.

Aude Oliva. 2005. CHAPTER 41 - gist of the scene. In Laurent Itti, Geraint Rees, and John K. Tsotsos, editors, *Neurobiology of Attention*, pages 251–256. Academic Press.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer.

James Pustejovsky. 2017. *ISO-Space: Annotating Static and Dynamic Spatial Information*, pages 989–1024. Springer Netherlands; Dordrecht.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. SemEval-2015 task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado. Association for Computational Linguistics.

Nils Reiter. 2015. Towards annotating narrative segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 34–38, Beijing, China. Association for Computational Linguistics.

Nils Reiter. 2020. Anleitung zur erstellung von annotationsrichtlinien. In *Anleitung zur Erstellung von Annotationsrichtlinien*, pages 193–202. De Gruyter.

Nils Reiter, Judith Sieker, Svenja Guhr, Evelyn Gius, and Sina Zarrieß. 2022. Exploring text recombination for automatic narrative level detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3346–3353, Marseille, France. European Language Resources Association.

Nils Reiter, Marcus Willand, and Evelyn Gius. 2019. A shared task for the digital humanities chapter 1: Introduction to annotation, narrative levels and shared tasks. 4(3).

Ruth Ronen. 1986. Space in fiction. 7(3):421–438. Publisher: Duke University Press, Porter Institute for Poetics and Semiotics.

Ruth Ronen. 1997. Description, narrative and representation. 5(3):274–286. Publisher: Ohio State University Press.

Marie-Laure Ryan. 2012. Space.

Marie-Laure Ryan, Kenneth Foote, and Maoz Azaryahu. 2016. *Narrating Space / Spatializing Narrative: Where Narrative Theory and Geography Meet*. Ohio State University Press.

Mareike Schumacher. 2023. *Orte und Räume im Roman: Ein Beitrag zur digitalen Literaturwissenschaft*. Digitale Literaturwissenschaft. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Sandeep Soni, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11723–11736, Toronto, Canada. Association for Computational Linguistics.

Gerard Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identiication. From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research (CELCR)*. John Benjamins Publishing Company.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Melissa Le-Hoa Võ, Sage EP Boettcher, and Dejan Draschkow. 2019. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. 29:205–210.

Melissa Le-Hoa Võ and Jeremy M. Wolfe. 2013. Differential electrophysiological signatures of semantic and syntactic scene processing. 24(9):1816–1823. Publisher: SAGE Publications Inc.

Wikimedia Foundation Inc. 2025. Wikivoyage – freie reiseinformationen rund um die welt.

Werner Wolf. 2007. Description as a transmedial mode of representation. general features and possibilities of realization in painting, fiction and music. In Werner Wolf and Walter Bernhart, editors, *Description in Literature and Other Media*, volume 2 of *Studies in Intermediality (SIM)*, pages 1–87. Brill.

Jeremy M. Wolfe, Melissa Le-Hoa Võ, Karla K. Evans, and Michelle R. Greene. 2011. Visual search in scenes involves selective and nonselective pathways. 15(2):77–84.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Albin Zehe, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer. 2021. Detecting scenes in fiction: A new segmentation task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3167–3177, Online. Association for Computational Linguistics.

Gabriel Zoran. 1984. Towards a theory of space in narrative. 5(2):309–335. Publisher: Duke University Press, Porter Institute for Poetics and Semiotics.

A Spatial Entities in the Corpora

In Table 4, we report the most frequent spatial entities in the two corpora.

B Example Sentences

In Table 5, we list a selection of sentences from the two corpora that do not unambiguously describe spatial frames.

C Prompts

In this section, we report the prompt variants in our experiment (§ 6). Based on the annotation guidelines, we formulate four different prompts as reported below.

C.1 EN-short

```
Your goal is to decide whether a sentence is a SPATIAL DESCRIPTION or not.

You will be provided with a sentence. You will answer with YES if that sentence is a SPATIAL DESCRIPTION. Otherwise, you will answer with NO.

In a SPATIAL DESCRIPTION, sensory features of spatial entities are described. These spatial entities form a static scene.
```

C.2 EN-medium

```
Your goal is to decide whether a sentence is a SPATIAL DESCRIPTION or not.

You will be provided with a sentence. You will answer with YES if that sentence is a SPATIAL DESCRIPTION. Otherwise, you will answer with NO.

A SPATIAL DESCRIPTION must meet all of the following criteria:

1. There is a description of a scene that consists of multiple entities.

2. The scene is static, it does not change.

3. There are descriptions of features that can be seen, felt, heard or smelled.
```

KOLIMO			Wikivoyage		
Entity	**Translation**	**Count**	**Entity**	**Translation**	**Count**
Stadt	City/Town	48003	Zimmer	Room	51408
Hafen	Port	16829	Stadt	City	45505
Museum	Museum	12975	Tür	Door	45287
Bahnhof	Station	11966	Fenster	Window	36323
Insel	Island	11777	Straße	Street/Road	35709
Park	Park	15051	Berg	Mountain	33416
Straße	Street/Road	20943	Tisch	Table/Desk	32672
See	Lake	12603	Platz	Place	31033
Platz	Place	13340	Erde	Earth	26549
Berg	Mountain	21811	Bett	Bed	21246

Table 4: The most frequent spatial entities in the two corpora according to the spatial entities collection by Herrmann et al. (2022). We also considered compounds and inflected forms of the reported lemmas.

4. The focus is on descriptions, not actions.

C.3 EN-long (KOLIMO)

Your goal is to decide whether a sentence is a SPATIAL DESCRIPTION or not.

You will be provided with a sentence. You will answer with YES if that sentence is a SPATIAL DESCRIPTION. Otherwise, you will answer with NO.

A SPATIAL DESCRIPTION must meet all of the following criteria:

- Space which can be described is the immediate environment where events could take place (at least theoretically), will take place in the future or have taken place in the past

- There are descriptive elements, not just the mere mention of space

- Scenes (arrangements of objects, background and foreground which are at least implicit) are described, not just a single object

- No unresolved references—what is described is always unambiguous

- There is no action, except for action that is expressed by verbs of perception and is related to space (see, hear ...)

- Generic, repeated actions can be part of a spatial description (e.g. sunset)

- Weather (rainfall, wind, clouds), daylight (solar altitude, dusk and dawn), ocean movements (waves, tide) and light (natural or artificial) are part of spatial descriptions, unless they explicitly take place suddenly or are part of individual actions

- The described space is static, stable and does not change during the description

- The described space is tangible (real, fictional, imagined, remembered, fantastic or dreamed), but not exclusively metaphorical or an abstraction

- The described qualities include all senses and are not limited to the visual

- Only complete descriptions are relevant, even if many sentences contain descriptive elements among

	Sentence	Translation	Source
1	Die Stadt erscheint mir kalt und fremd und widert mich.	The city seems cold and foreign to me and disgusts me.	Felix Hollaender: Die Briefe des Fräulein Brandt (1918)
2	Rings auf der bloßen Erde lagen lauter Schläfer.	All around on the bare earth were lying many sleepers.	Jakob Wassermann: Alexander in Babylon (1905)
3	Auch ist drinnen kein Platz mehr.	There is no room left inside either.	Fritz Mauthner: Der neue Ahasver (1882)
4	Die Kleinstadt Adorf liegt im Vogtlandkreis am Nordrand des Elstergebirges.	The small town of Adorf is located in Vogtlandkreis on the northern edge of the Elster mountains.	Wikivoyage: Adorf
5	Katharinenkapelle: Die Kapelle steht auf dem 493 m hohen Katharinenberg, es ist der zweithöchste Berg des Kaiserstuhls.	Katharinenkapelle: The chapel stands on the 493 meters high Katharinenberg, it is the second highest mountain in the Kaiserstuhl.	Wikivoyage: Endingen am Kaiserstuhl
6	Vorbei am Balcon du Ranc pointu fällt die Straße nun ab um die ersten Häuser und Campingplätze von Saint-Martin-d'Ardèche zu erreichen [sic].	Passing the Balcon du Ranc pointu, the road now descends to reach the first houses and campsites of Saint-Martin-d'Ardèche.	Wikivoyage: Gorges de l'Ardèche
7	Neben den Badestränden kann man auf den Cerro La Cruz laufen, einem etwa 1000 m hohen Berg, auf dem sich ein großes Kreuz befindet (ca. 30-45 min Fußmarsch je nach Kondition).	In addition to the beaches, you can walk up the Cerro La Cruz, a mountain about 1000 meters high, on which there is a large cross (approx. 30-45 min walk depending on fitness level).	Wikivoyage: Via Carlos Paz
8	Girlanden mit Lampions quer über den Hof von Flurfenster zu Flurfenster.	Garlands with lanterns across the courtyard from corridor window to corridor window.	Hans Ostwald: Das Zillebuch (1929)
9	Delaware Park: Größter Park in Buffalo mit gepflegten Grünflächen und einem See.	Delaware Park: Largest Park in Buffalo with well-tended green spaces and a lake.	Wikivoyage: Buffalo/Norden
10	Vor mir wachsen die geheimnisvollen, glutroten Korallen aus der Tiefe des Wassers, sie breiten ihr mystisches Geäst aus über den Himmel, sie flechten ein Netz durch Luft und Wolken, ein Netz von blutfarbenen Zweigen, an dem weiße Perlen schimmern.	In front of me, the mysterious, glowing red corals grow from the depths of the water, spreading their mystical branches across the sky, weaving a net through the air and clouds, a net of blood-colored branches on which white pearls shimmer.	Nataly von Eschstruth: Die Bären von Hohen-Esp (1922)

	Sentence	Translation	Source
11	Auch hatte sie hier den Apparat dicht neben sich, während das andere Telephon sich im Bibliothekzimmer befindet.	She also had the device [or *phone*] right next to her, while the other phone was in the library room.	Hugo Bettauer: Die freudlose Gasse (1924)
12	Ein Wachtmantel von gelbem Tuch mit grünem Kragen – grün und gelb waren die Farben der Stadt – hing am Nagel, ein Bauer mit einem bunten, klugen Zeisig von der Decke.	A watchman's coat of yellow cloth with a green collar—green and yellow were the colors of the city—hung from the nail, a cage [or *peasant*] with a colorful, clever siskin from the ceiling.	Wilhelm Raabe: Das letzte Recht (1910)

Table 5: Examples for annotated sentences.

others

- The sentences are complete and in German

C.4 EN-long (Wikivoyage)

Your goal is to decide whether a sentence is a SPATIAL DESCRIPTION or not.

You will be provided with a sentence. You will answer with YES if that sentence is a SPATIAL DESCRIPTION. Otherwise, you will answer with NO.

A SPATIAL DESCRIPTION must meet all of the following criteria:

- Space which can be described is the immediate environment where events could take place (at least theoretically), will take place in the future or have taken place in the past

- There are descriptive elements, not just the mere mention of space

- Scenes (arrangements of objects, background and foreground which are at least implicit) are described, not just a single object

- No unresolved references: what is described is always unambiguous

- There is no action, except for action that is expressed by verbs of perception and is related to space (see, hear ...)

- Generic, repeated actions can be part of a spatial description (e.g. sunset)

- Weather (rainfall, wind, clouds), daylight (solar altitude, dusk and dawn), ocean movements (waves, tide) and light (natural or artificial) are part of spatial descriptions, unless they explicitly take place suddenly or are part of individual actions

- The described space is static, stable and does not change during the description

- The described space is tangible (real, fictional, imagined, remembered, fantastic or dreamed), but not exclusively metaphorical or an abstraction

- The described qualities include all senses and are not limited to the visual

- No route descriptions from A to B

- The geographical location of a named entity is not a spatial description

- Only complete descriptions are

relevant, even if many sentences contain descriptive elements among others

- The sentences are complete and in German

C.5 DE-long (KOLIMO)

Du sollst entscheiden, ob ein Satz eine RAUMBESCHREIBUNG ist oder nicht.

Du bekommst einen Satz, und du wirst mit JA antworten, falls dieser Satz eine RAUMBESCHREIBUNG ist. Ansonsten wirst du mit NEIN antworten.

Eine RAUMBESCHREIBUNG muss alle folgenden Kriterien erfüllen:

- Raum, der beschrieben werden kann, ist die unmittelbare Umgebung, in der das Geschehen (zumindest theoretisch) stattfinden könnte, in der Zukunft stattfinden wird oder in der Vergangenheit stattgefunden hat

- Es gibt beschreibende Elemente, nicht die bloße Nennung von Raum

- Es werden Szenen (zumindest implizite Arrangements von Objekten, Hintergrund und Vordergrund) beschrieben, nicht nur ein einzelnes Objekt

- Keine unaufgelösten Referenzen – es ist immer eindeutig, was beschrieben wird

- Es gibt keine Handlung, außer solche, die durch Verben der Wahrnehmung ausgedrückt wird und sich auf den Raum bezieht (sehen, hören . . .)

- Generische, wiederholte Handlungen können Teil einer Raumbeschreibung sein (z.B. das Untergehen der Sonne)

- Wetter (Niederschlag, Wind, Wolken), Tageslichtphasen (Sonnenstand, Dämmerung), Meeresbewegungen (Wellen, Gezeiten) und Licht (von Lampen oder der Sonne) sind Teil von Raumbeschreibungen, solang sie nicht explizit plötzlich und in individuellen Handlungen vorkommen

- Der beschriebene Raum ist statisch, stabil und verändert sich nicht während der Beschreibung

- Der beschriebene Raum ist konkret (real, fiktional, imaginiert, erinnert, phantastisch, geträumt), aber nicht ausschließlich metaphorisch oder eine Abstraktion

- Die beschriebenen Qualitäten umfassen alle Sinne und sind nicht auf das Visuelle beschränkt

- Nur vollständige Beschreibungen sind relevant, auch wenn viele Sätze unter anderem raumbeschreibende Elemente enthalten

- Die Sätze sind vollständig und auf Deutsch

C.6 DE-long (Wikivoyage)

Du sollst entscheiden, ob ein Satz eine RAUMBESCHREIBUNG ist oder nicht.

Du bekommst einen Satz, und du wirst mit JA antworten, falls dieser Satz eine RAUMBESCHREIBUNG ist. Ansonsten wirst du mit NEIN antworten.

Eine RAUMBESCHREIBUNG muss alle folgenden Kriterien erfüllen:

- Raum, der beschrieben werden kann, ist die unmittelbare Umgebung, in der das Geschehen (zumindest theoretisch) stattfinden könnte, in der Zukunft stattfinden wird oder in der Vergangenheit stattgefunden hat

- Es gibt beschreibende Elemente, nicht die bloße Nennung von Raum

- Es werden Szenen (zumindest implizite Arrangements von Objekten, Hintergrund und Vordergrund) beschrieben, nicht nur ein einzelnes Objekt

- Keine unaufgelösten Referenzen – es ist immer eindeutig, was beschrieben wird

- Es gibt keine Handlung, außer solche, die durch Verben der Wahrnehmung ausgedrückt wird und sich auf den Raum bezieht (sehen, hören . . .)

- Generische, wiederholte Handlungen können Teil einer Raumbeschreibung sein (z.B. das Untergehen der Sonne)

- Wetter (Niederschlag, Wind, Wolken), Tageslichtphasen (Sonnenstand, Dämmerung), Meeresbewegungen (Wellen, Gezeiten) und Licht (von Lampen oder der Sonne) sind Teil von Raumbeschreibungen, solang sie nicht explizit plötzlich und in individuellen Handlungen vorkommen

- Der beschriebene Raum ist statisch, stabil und verändert sich nicht während der Beschreibung

- Der beschriebene Raum ist konkret (real, fiktional, imaginiert, erinnert, phantastisch, geträumt), aber nicht ausschließlich metaphorisch oder eine Abstraktion

- Die beschriebenen Qualitäten umfassen alle Sinne und sind nicht auf das Visuelle beschränkt

- Keine Streckenbeschreibungen von A nach B

- Die geographische Lage einer benannten Entität ist keine Raumbeschreibung

- Nur vollständige Beschreibungen sind relevant, auch wenn viele Sätze unter anderem raumbeschreibende Elemente enthalten

- Die Sätze sind vollständig und auf Deutsch

D LLMs Prompting Experiment Settings

We run all the open-source model experiments using their 8-bit quantization versions via the HuggingFace transformers library. We use a single NVIDIA RTX A6000 GPU to run all our open-source experiments, while we call OpenAI's API for the GPT-4o experiments. We set the LLMs' generation temperature to zero at all our prompting calls, and we set the seed to 42 whenever possible, to allow for reproducibility.

E Evaluation of LLMs Annotations

We report the results for our 28 model-prompt variants in this section. Table 6 shows the results of GPT-4o prompt variants, while the results of the open-source model-prompt variants are reported in Table 7.

		Literary Dataset					Non-Literary Dataset				
Model	**Prompt**	**Acc.**	**P**	**R**	**F1**	**Rat.**	**Acc.**	**P**	**R**	**F1**	**Rat.**
GPT-4o	EN-short	.87	.38	**.81**	.51	.18	.72	.38	**.70**	.50	.37
	EN-med	.93	.57	.55	.56	**.08**	.85	.67	.43	**.53**	.13
	EN-long	.94	**.64**	.63	**.64**	**.08**	.84	**.97**	.19	.32	.04
	DE-long	.93	.58	.69	.63	.10	.82	.76	.16	.27	.04

Table 6: GPT-4o Results.

			Literary Dataset					Non-Literary Dataset				
Family	**Size**	**Prompt**	**Acc.**	**P**	**R**	**F1**	**Rat.**	**Acc.**	**P**	**R**	**F1**	**Rat.**
Gemma2	2B	EN-short	.64	.17	.88	.29	.43	.46	.26	.96	.41	.72
		EN-med	.82	.29	.82	.43	.24	.61	.32	.83	.46	.52
		EN-long	.60	.17	**.96**	.29	.47	.56	.30	**.97**	.46	.63
		DE-long	.76	.24	.84	.37	.30	.78	.47	.58	.52	.25
	9B	EN-short	.53	.14	.90	.24	.54	.57	.29	.77	.42	.54
		EN-med	.81	.30	.89	.44	.26	.72	.39	.75	.51	.38
		EN-long	.78	.26	.89	.41	.29	.83	.60	.43	.50	.14
		DE-long	.82	.31	.88	.45	.24	.84	.87	.21	.34	.05
	27B	EN-short	.60	.16	.91	.28	.47	.60	.30	.75	.43	.50
		EN-med	.80	.28	.88	.43	.27	.73	.40	.72	.52	.36
		EN-long	.68	.20	.95	.33	.40	.83	.56	.62	**.59**	.22
		DE-long	.86	.37	.86	.52	.20	.84	.81	.26	.40	.06
Qwen2.5	3B	EN-short	.92	.00	.00	.00	.00	.80	.00	.00	.00	.00
		EN-med	.92	.00	.00	.00	.00	.80	.00	.00	.00	.00
		EN-long	.92	.00	.00	.00	.00	.80	.00	.00	.00	.00
		DE-long	.92	.00	.00	.00	.00	.80	.00	.00	.00	.00
	7B	EN-short	.85	.31	.60	.41	.17	.66	.31	.57	.40	.37
		EN-med	.93	.56	.57	.57	**.09**	.76	.40	.42	.41	**.21**
		EN-long	.83	.30	.77	.43	.22	.82	.58	.35	.44	.12
		DE-long	.87	.36	.70	.47	.17	.82	.80	.10	.18	.02
	32B	EN-short	.92	.50	.73	.59	.12	.71	.36	.61	.46	.33
		EN-med	.94	.63	.65	.64	**.09**	.81	.54	.38	.45	.14
		EN-long	**.95**	**.83**	.56	**.67**	.06	.81	**1.0**	.06	.12	.01
		DE-long	.94	.62	.71	.66	.10	.81	**1.0**	.06	.12	.01

Table 7: Open-source models Results.