Recognising non-named spatial entities in literary texts: a novel spatial entities classifier⋆

Daniel Kababgi[1], Gulia Grisot[2], Federico Pennino[3] and Berenike Herrmann[1]

[1]Universität Bielefeld
[2]University of Cambridge
[3]Università di Bologna

Abstract
Predicting spatial representations in literature is a challenging task that requires advanced machine learning methods and manual annotations. In this paper, we present a study that leverages manual annotations and a BERT language model to automatically detect and recognise non-named spatial entities in a historical corpus of Swiss novels. The annotated data, consisting of Swiss narrative texts in German from the period of 1840 to 1950, was used to train the machine learning model and fine-tune a deep learning model specifically for literary German. The annotation process, facilitated by the use of Prodigy, enabled iterative improvement of the model's predictions by selecting informative instances from the unlabelled data. Our evaluation metrics (F1 score) demonstrate the model's ability to predict various categories of spatial entities in our corpus. This new method enables researchers to explore spatial representations in literary text, contributing both to digital humanities and literary studies. While our study shows promising results, we acknowledge challenges such as representativeness of the annotated data, biases in manual annotations, and domain-specific language. By addressing these limitations and discussing the implications of our findings, we provide a foundation for future research in sentiment and spatial analysis in literature. Our findings not only contribute to the understanding of literary narratives but also demonstrate the potential of automated spatial analysis in historical and literary research.

1. Introduction

Building on previous work examining fictional space and sentiment in Swiss-German narrative [7], this paper reports on the development and evaluation of a novel machine learning model for the analysis of fictional space in literary texts.[1]

In recent criticism across disciplines there has been an increased emphasis towards considering place and space as crucial factors in understanding social, cultural, and historical phenomena. This perspective on spatiality is generally referred to as the 'spatial turn' [9, 10, 19], and, in literary studies, it highlights the integral components of space representation in how we under-

---

[1]The dataset and the code can be found here: https://github.com/DanielKababgi/spatial_entities_classifier

stand and contextualise narrative and fictional texts. The exploration of spatial representations in literary works offers valuable insights into the landscapes, the constructed environments and their social implications within narratives, as well as into the cultural and socio-political constructs surrounding certain images. while there are several valuable proposals to a quantitative approach to spatial research [24, 2], including for example a differentiation of space as background and place more specifically as locus of events [23, 15], we however set a different focus.

In this paper, we present a case study on the prediction of what we call 'non-named spatial entities' (NNSE) in a historical corpus of Swiss-German novels using a deep learning model in conjunction with BERT and Prodigy. By combining manual annotations and advanced machine learning methods, we aim to automatically detect and recognise NNSE within the literary narratives via a similar approach to named entity recognition (NER).

NER techniques are used to identify and categorise text segments that refer to entities such as people, places, or companies, and that 'constitute proper names' [11]. The latest NER techniques rely on manually annotated text corpora, which are automatically analysed to build models that capture language use and grammar. These models can then identify and classify entities in new, unprocessed documents. State-of-the-art NER systems come with pre-built models trained on extensive collections of annotated documents, like news articles. These models typically perform well and are ideal for specific applications such as analysing customer feedback or extracting locations and characters. However, when applied to documents with linguistic features not well-represented in the training data, such as literary texts, NER performance can decline, increasing the likelihood of errors, also increased for most languages other than English.

Within literary studies, various scholars have used NER, particularly to identify fictional characters [20] , build social networks [5], identify geographical locations [3], to assign headings to novels [16], or to analyse relationships between literary works [21, 22, 17]. Only few researchers until recently however focused on the identification of NNSE, i.e. those terms, or elements of space representation, which are not necessarily named geographical locations, like Berlin, London, or Zurich, and that therefore typically cannot be located on a map. It is this kind of entity that generally contributes most to effectively to create the so-called 'storyworld' [18]: simple terms or phrases such as 'mountain', 'bridge', 'beach' or 'cave', as well as objects and architectural parts that make them tangible, such as 'window', 'table', or 'wall'.

A similar perspective has been considered by Schumacher, Flüh, and Nantke [14], who used conditional random fields to automatically annotate among other things non-named places, which is conceptually similar to NNSEs. However, the operationalization of space in their research is focused on places that can be found on a map, leaving therefore the broader concept of NNSEs unexplored. Also, the popular BookNLP toolkit by Bamman is able to work to identify what they call 'locations' (for natural entities) and facilities' (for man-made structures) in English-language texts with an accuracy of up to 90%[2]. While BookNLP is able to differentiate NNSEs somewhat according to our own needs, we propose a distinction for 'facilities' into more discrete classes shown in Section 2.1. This paper sets out to fill this gap, training a model that will help us identify spatial elements in narrative.

---

[2]As is shown on their official GitHub page: https://github.com/booknlp/booknlp?tab=readme-ov-file

Table 1
Examples of words for each category of NNSE

Category	Exampels
interior	*Abstellkammer* (storage room), *Wohnzimmer* (living room), *Küche* (kitchen)
urban	*Bibliothek* (library), *Kloster* (abby), *Vorstadt* (suburb)
rural	*Bauernhaus* (farmhouse), *Garten* (garden), *Schweinestall* (pigsty)
natural	*Berg* (mountain), *Fluss* (river), *Wald* (forest)

2. Method

2.1. Spatial categories

In order to train our model to recognise literary space, we decided to train it not only to be able to recognise non-named spatial entities (NNSEs), but also to distinguish among four different types of spatial environments. We decided to base our categorisation on the research by Grisot and Herrmann [7], who looked at the sentiment encoded in narrative texts in relation to both named and non-named entities. They used a dictionary based approach, collecting spatial terms for geographical locations as well as for non-named entities, distinguishing in particular the categories 'rural', 'natural' and 'urban'. While these three categories offered a promising base, we felt that for a more comprehensive perspective on the spaces and places rendered in fictional texts we also needed to include spatial elements describing the interiors/indoor space of buildings and rooms. We therefore created an additional category of NNSE, 'interior'. Some examples for each category are shown in Table 1 above.

2.2. Annotations and model training

To produce the training set, two annotators were provided with written guidelines and trained in person to understand the difference between the various NNSE categories. They were then instructed to use the platform Prodigy [12], which allowed them to read sentences from the training set in random order, and to annotate NNSEs directly on the interface by adding labels to individual tokens. For the annotation process, six novels were sampled from the complete corpus of Swiss-German novels [8].[3] The novels were then split into sentences (N=9,062), which were manually annotated by the annotators. The annotators featured a high inter-annotator agreement (Cohen's Kappa) for the NNSE types 'interior', 'natural', and 'rural', and medium agreement for 'urban'[4], as well as a high agreement for the distinction of NNSEs against a not-NNSE token (Cohen's Kappa = 0.898).

The values for annotators' agreement in relation to individual types are shown in Table 2. Also shown in Table 2 are the number of sentences in our dataset in which the various classes of spatial entities were identified by manual annotation. Type O shows the number of sentences where no NNSEs were identified in the annotation process. These amount to over 86% of the

---

[3]These are: *Der Wetterwart* (1905) by Jacob Christoph Heer, *Heimatscholle* (1914) by Maria Goswina von Berlepsch, *Berge und Menschen* (1911) by Heinrich Federer, *Heidis Lehr- und Wanderjahre* (1880) by Johanna Spyri, *Friedli, der Kolderi* (1891) by Carl Spitteler, and *Martin Salander* (1886) by Gottfried Keller.

[4]The low score for 'urban' is explainable by the low number of occurrencesof NNSE of this type in the dataset.

Table 2

Types of NNSE in complete dataset with inter-annotator aggreement (Cohen's Kappa) between the two annotators for each class, number of sentences in the complete dataset in which the types of NNSE were identified by manual annotation. Class O marks sentences with no NNSE.

Class	Cohen's Kappa	n Occurrences	Percentage
interior	0.933	412	4.5
urban	0.608	196	2.2
rural	0.775	315	3.5
natural	0.857	328	3.6
O	0.896	7811	86.2

dataset, making the distribution of the five categories in our dataset unbalanced. However, the NNSE categories are much closer together, ranging between 2.2% (urban) and 4.5% (interior) of all sentences.

After the annotation process, sentences were randomly assigned to either the training dataset (80%, N=7,249) or the test dataset (20%, N=1,813). This was done according to common best practices for training a machine learning model [6]. To train the classifier, PyTorch version 2.1.1 was used as the deep learning framework [13]. In conjunction with PyTorch, the popular hugging face library (version 4.35.2) was used to load and interact with the language model [25].

The model *gbert-large* by *deepset* was utilised as input layer for the token classifier, since it outperformed all other language models [4]. The model classifies each token of a given text, and attempts to predict whether the token under consideration can be classified as a NNSE (one of the four types mentioned above) or whether it can be considered not a spatial term (O). It was tested if the model performs better with the complete, unbalanced training dataset (N=7,249 sentences) or with a more balanced, downsampled training dataset (N=2,004 sentences). The downsampled training dataset was composed of sentences including at least one NNSE (N=1,002) and a random selection of sentences with no NNSE of equal size. The downsampled train dataset was then split again into a final training dataset (N=1,603) and a validation set (N=401).

The training was repeated for 17 epochs with a learning rate of 5e-6 and a dropout of 0.1. These parameters are determined by extensive hyperparameter testing. Since the maximum length of all sentences is 40 words or 61 tokens, the max length for the BERT model was set to 64 tokens.

3. Results

With the annotation and training process described above, we produced a classifier able to 1) identify NNSEs in a given sentence, and 2) classify the identified NNSE as belonging to one of the four discrete classes: 'rural', 'urban', 'natural' or 'interior'.[5] To find the best parameters,

---

[5]The results reported here should be understood as a report of an ongoing process, rather than as a final product. For example, a theoretically more sound model may understand 'interior' not as a category of its own, but allow
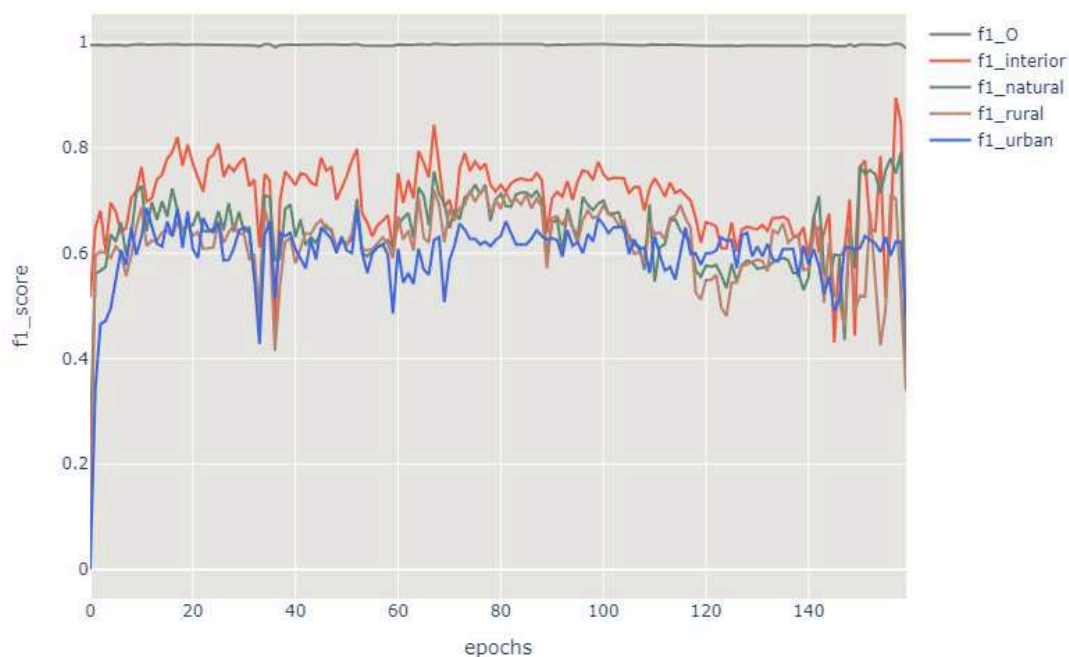
Fig 1: F1 scores for all five classes

Figure 1: F1 scores on the validation dataset for all classes, including O, over 160 epochs. Highest scores for all trained epochs are by class O, the lowest scores are for identifying urban NNSE

the F1 score for each class was calculated for validation after each epoch. While the F1-score for the O class was very high for every epoch, the scores for the discrete categories of NNSE fluctuated quite strongly.

Figure 1 illustrates the performance evaluation of the classifier on the validation dataset across different epochs. Class O, which represents tokens not classified as NNSE according to our guidelines, consistently achieves an almost perfect F1 score of 1. However, after 17 epochs, there is a significant decline in performance. Among the other classes, the 'interior' class performs best with an F1 score of 0.792, while the 'urban' class performs worst with an F1 score of 0.632 on the validation dataset. For a more detailed analysis, Figure 2 presents the average F1 scores for the 'interior', 'urban', 'rural', and 'natural' classes on the validation dataset, excluding class O. The black dotted line indicates the highest overall F1 score of 0.743.

Table 3 below presents the final scores on the test dataset. Class O, with an F1 score of 0.99, significantly outperforms all other classes, as is expected. The 'interior' class follows with an F1 score of 0.60. The remaining classes have F1 scores ranging between 0.64 and 0.53.

---

for categorization into interior-rural or interior-urban. We are planning to run a set of annotations on the interior items to add this type of information and explore the differences.
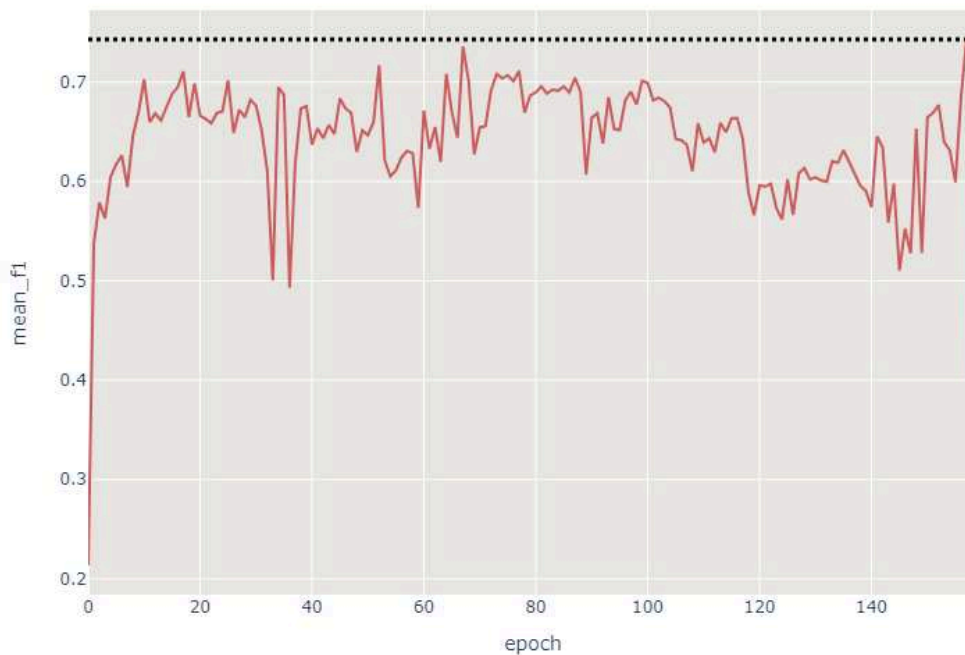
Fig 2: Mean f1 score for types of NNSE per epoch

Figure 2: Mean F1 score for the classes interior, urban, rural, and natural over 160 epochs. The dotted black line indicates the highest mean F1-score of 0.743.

Table 3
F1 scores per class on the test dataset. After class O the second best score is for the class interior. The worst performance is for class rural.

Class	F1 score	Precision	Recall
interior	0.6079	0.4673	0.8696
urban	0.5333	0.4034	0.7869
rural	0.5573	0.4620	0.7025
natural	0.6468	0.5353	0.8125
O	0.9984	0.9996	0.9972

In addition to the general F1 scores for each class, we analysed false classifications across all five classes. The separation between NNSEs and class O was highly effective, with tokens belonging to class O rarely being incorrectly classified as NNSE. Conversely, the most common error for all types of spatial entities was their misclassification as class O. Notably, the differentiation between various spatial entities was generally accurate, except for the 'urban' class, which was occasionally misclassified as 'rural'.
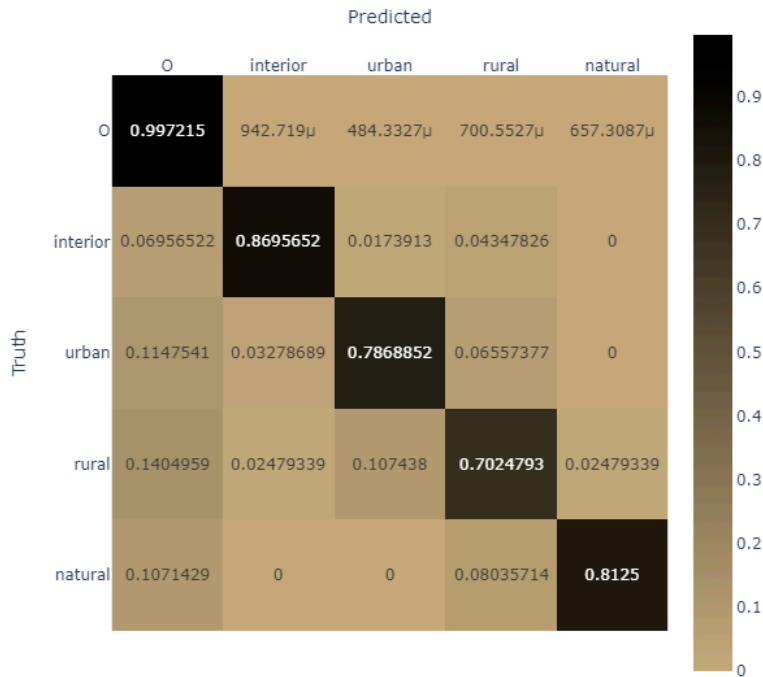
Fig 3: Confusion matrix, NNSE classifer

Figure 3: Confusion matrix for all five classes. All classes are classified correctly as themselves, while the biggest error is not being recognized as a NNSE. The biggest error after that is by misclassifing 'rural' as 'urban' 13% error rate for being classified as 'rural'.

Figure 3 displays a confusion matrix for all classes. The highest error rate for all four NNSE types involves their misclassification as class O, ranging from 6.9% for 'interior' to 14% for 'rural'. The biggest error after that is by misclassifing 'rural' as 'urban' 13% error rate for being classified as 'rural'.

4. Conclusion and further Research

In this work, we have developed a tool that already in its present state facilitates valuable quantitative spatial research on 19th and early 20th century German-language literary corpora. While out-of-the-box solutions typically only provide Named Entity Recognition (NER) models, to the best of the authors' knowledge, a classification of non-named spatial entities as conducted here, classifying each entity into different types, has never been published before for German. Schumacher, Flüh, and Nantke [14] developed a classifier based on Conditional Random Fields (CRF), which includes several more categories than our classifier, but it can only detect places, not their types. Bamman's *bookNLP* is able to differentiate between *locations*,

which alignes with our 'natural' type, and *facilities*, which covers 'urban' and 'rural' spatial entities. For comparison, we also tested the large language model Llama3 7b [1] with a few-shot prompt for recognising unnamed spatial entities (NNSE). The automatic classification on the test dataset with Llama3 resulted in only a 5.6% partial match with the manual annotation, and a 0.7% perfect match. The main issue was the model's tendency to hallucinate new NNSEs when attempting to continue a sentence, contrary to instructions.

The high performance in distinguishing spatial entities from non-spatial tokens is unsurprising, as this was the least contentious aspect during the evaluation of the annotation process. The high error rate of 'rural' being misclassified as 'urban' but not vice versa can be explained by the prevalence of 'rural' space in the training data. Additionally, the boundary between 'rural' and 'urban', as described in the guidelines, is more 'fuzzy' compared to the respective distinctions to 'interior' and 'natural'. This fuzziness may be aggravated by the inherent ambiguity of using sentences as training units.

This classifier is considered a work in progress, as it has currently been exclusively trained on Swiss-German texts from the late 19th to early 20th century. Potential improvements include gathering more training data and adapting the *gbert-large* model to Swiss-German literary texts from the long 19th century and beyond, as well as remodelling the categories to include interior-urban and interior-rural.

We plan to utilise this classifier to explore the remainder of the Swiss-German novel corpus built by Herrmann and Grisot [8], qualitatively examining patterns in the representation of space, with a particular focus on interior items. Subsequently, we intend to extend our research to a broader corpus of German literature. Methodologically, we plan to evaluate the use of synthetic training data provided by generative AI to enhance our model. One key aspect of space that will be analysed in-depth in future research is the relationship between affect and space in literature, building upon the previous work of Grisot and Herrmann [7].

References

[1] AiMeta. *Llama 3 Model Card*. 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL%5C%5FCARD.md.

[2] F. Barth. "Konzept und Klassifikation literarischer Raumentitäten". In: *Informatik 2020* (2021), pp. 1281–1293. DOI: 10.18420/inf2020_120.

[3] S. Bushell, J. O. Butler, D. Hay, and R. Hutcheon. "Digital Literary Mapping: I. Visualizing and Reading Graph Topologies as Maps for Literature". In: *Cartographica: The International Journal for Geographic Information and Geovisualization* 57.1 (2022), pp. 11–36. DOI: 10.3138/cart-2021-0008.

[4] B. Chan, S. Schweter, and T. Möller. *German's Next Language Model*. 2020. arXiv: 2010.10906 [cs]. URL: http://arxiv.org/abs/2010.10906. Pre-published.

[5] N. Dekker, T. Kuhn, and M. Van Erp. "Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels". In: *PeerJ Computer Science* 5 (2019), e189. DOI: 10.7717/peerj-cs.189.

[6] A. Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Second edition. Covid-19 Collection. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2019. 819 pp.

[7] G. Grisot and B. Herrmann. "Examining the Representation of Landscape and Its Emotional Value in German-Swiss Fiction between 1840 and 1940". In: *Journal of Cultural Analytics* 8.1 (2023). DOI: 10.22148/001c.84475.

[8] G. Grisot and B. Herrmann. *Swiss German Novel Collection (ELTeC-gsw)*. Version 2.0.0. Zenodo, 2021. DOI: 10.5281/zenodo.4584544.

[9] W. Hallet and B. Neumann. *Raum Und Bewegung in der Literatur: Die Literaturwissenschaften Und Der Spatial Turn*. Bielefeld: transcript Verlag, 2015.

[10] E. W. B. Hess-Lüttich. "Spatial Turn: On the Concept of Space in Cultural Geography and Literary Theory". In: *Meta-Carto-Semiotics* 5.1 (2017), pp. 27–37.

[11] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. 2024.

[12] I. Montani and M. Honnibal. *Prodigy: A Modern and Scriptable Annotation Tool for Creating Training Data for Machine Learning Models*. Explosion, 2018. URL: https://prodi.gy/.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[14] M. Schumacher, M. Flüh, and J. Nantke. "Place and Space in Literature Named Entity Recognition as a Possibility for Spatial Modelling in Computational Literary Studies". In: *Geographical Research in the Digital Humanities*. Ed. by F. Dammann and D. Kremer. transcript Verlag, 2024, pp. 83–112. DOI: 10.14361/9783839469187-006.

[15] M. K. Schumacher. *Orte und Räume in Romanen*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023. 234 pp. DOI: 10.1007/978-3-662-66035-5_1.

[16] M. Short. "Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels". In: *Cataloging & Classification Quarterly* 57.5 (2019), pp. 315–336. DOI: 10.1080/01639374.2019.1653413.

[17] R. T. Tally, ed. *Geocritical Explorations: Space, Place, and Mapping in Literary and Cultural Studies*. New York: Palgrave Macmillan US, 2011. DOI: 10.1057/9780230337930.

[18] R. T. Tally. "The Space of the Novel". In: *The Cambridge Companion to the Novel*. Ed. by E. Bulson. 1st ed. Cambridge University Press, 2018, pp. 152–167. DOI: 10.1017/9781316659694.011.

[19] R. T. Tally. *Topophrenia: Place, Narrative, and the Spatial Imagination*. Indiana University Press, 2019-01-02. DOI: 10.2307/j.ctv7r40df. JSTOR: 10.2307/j.ctv7r40df.

[20] H. Vala, D. Jurgens, A. Piper, and D. Ruths. "Mr. Bennet, His Coachman, and the Archbishop Walk into a Bar but Only One of Them Gets Recognized: On The Difficulty of Detecting Characters in Literary Texts". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 769–774. DOI: 10.18653/v1/D15-1088.

[21] K. Van Dalen-Oskam. "Names in Novels: An Experiment in Computational Stylistics". In: *Literary and Linguistic Computing* 28.2 (2013), pp. 359–370. DOI: 10.1093/llc/fqs007.

[22] K. Van Dalen-Oskam, M. Marx, I. Sijaranamual, K. Depuydt, B. Verheij, and V. Geirnaert. "Named Entity Recognition and Resolution for Literary Studies". In: *Computational Linguistics in the Netherlands Journal* 4 (2014), pp. 121–136. URL: http://www.clinjournal.org/node/62.

[23] G. Viehhauser. "Zur Erkennung von Raum in Narrativen Texten: Spatial Frames Und Raumsemantik Als Modelle Für Eine Digitale Narratologie Des Raums". In: *Reflektierte Algorithmische Textanalyse*. Ed. by N. Reiter, A. Pichler, and J. Kuhn. De Gruyter, 2020, pp. 373–388. DOI: 10.1515/9783110693973-015.

[24] G. Viehhauser-Mery and F. Barth. "Towards a Digital Narratology of Space". In: *Book of Abstracts DH2017*. Dh2017. Montreal, 2017. URL: https://dh2017.adho.org/abstracts/413/413.pdf.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. Version 5. 2019. DOI: 10.48550/arxiv.1910.03771. Pre-published.