Scalable Verb-Based Literary Semantics

Hans Ole Hatzel[1], Haimo Stiemer[2], Evelyn Gius[2], and Chris Biemann[1]

[1] Department of Informatics, University of Hamburg, Hamburg, Germany
[2] Institute of Linguistics and Literary Studies, TU Darmstadt, Darmstadt, Germany

Abstract

We present a novel, scalable approach for coarse-grained semantic profiling of literary texts, intended as an alternative to existing explorative approaches like topic modeling. We propose using verb classes, essentially a coarse-grained variant of word sense disambiguation, to build interpretable semantic text profiles. In our approach, we classify verbal phrases based on the semantic class of their full verb, as formalized in GermaNet and WordNet. Our 10,000 manual annotations serve as a testbed for our retrieval-augmented generation (RAG) approach to classifying semantic verb classes. We test foundation models in multiple prompting setups with few-shot learning and RAG, illustrating that our approaches can almost, but not quite, match the performance of human annotators. Our application to four genre corpora (Romantic fiction, Regional literature, Crime fiction, Adventure fiction) of German-language texts demonstrates that such semantic profiles can reveal the thematic focuses of document collections. We release our corpus of more than 10,000 verb class annotations.[1]

Keywords: verb classes, prompting, computational literary studies, large language model, rag, semantics, genre, themes

1 Introduction

What is a text about? At first glance, the question seems straightforward; yet the wide range of scholarly approaches contests this assumption. For example, non-computational, qualitative German literary studies in the field of thematology are dominated by research that focuses on concepts such as “Stoff” or “Motiv” [4; 7], which roughly means a canonical story pattern or plot material. However, their definition is often based on interpretative reading and lacks operational characteristics. In Computational Literary Studies (CLS), topic modeling has become an established technique for the exploration of text content [10]. Nevertheless, in addition to its dependence on hyperparameters for which ideal values cannot trivially be determined [20], it is not a straightforward approach for thematic analysis, as it requires an interpretation of the ‘topics’. Resources like WordNet [6] and its German-language counterpart GermaNet [9] do not depend on interpretation and parametrization because they provide a general semantic approach, listing numerous senses per lemma. FrameNet [1] defines a rich inventory of semantic frames and their roles. These fine-grained approaches, however, pose practical challenges. Relying on the occurrence statistics of hundreds of senses or frames across texts leads to sparsity issues, making direct comparison and semantic profiling difficult in practice. Our approach, following a previously published concept [11], aims to enhance both interpretability and generalizability by identifying only the broad semantic areas of specific actions, thereby gaining a preliminary overview of the thematic profile that scales from single literary texts to entire corpora. We achieve this by using not the synsets but the semantic verb classes or fields in GermaNet, which are based on the verb classes by Levin,

Hans Ole Hatzel, Haimo Stiemer, Evelyn Gius, and Chris Biemann. “Scalable Verb-Based Literary Semantics.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 450–462. https://doi.org/10.63744/IwD9hAGns9BK.

[1] The data is available at `https://github.com/forTEXT/verbclasses/`

instead [14].[2] They classify all verbs into 15 categories (see Table 5 for a complete list of all classes). Verbs are central carriers of action and process [16]; they can be said to represent the events in a text at a micro level. Aggregated verb analysis provides a profile-like representation of a given text's event type distribution, which in turn correlates with genres and thematic patterns [12]. Moreover, verb class analysis facilitates the interpretation of its results, since verb classes do not constitute an overly fine-grained categorical system.

In this paper, we first present the manual annotation task for verb classes (section 2) and explore the resulting annotation data (section 3). In the second step, we automate the annotation of these classes by comparing various prompting techniques across a range of large language models (LLMs) and validate our automated methods using manually annotated data (section 4). We then apply our approach to four genre corpora to examine the extent to which verb classes also indicate genres (section 5). Finally, we provide a preliminary evaluation of our approach and the results (section 6).

2 Annotation Procedure

	Number of Annotations				
Text Title	Annotator 1	Annotator 2	Annotator 3	Fleiss' Kappa	Tokens
Das Erdbeben in Chili	666	725	677	0.64	6499
Der blonde Eckbert	898	930	54	0.85	7508
Die Judenbuche	52	0	2269	0.73	20129
Die Verwandlung	2296	2283	53	0.69	22299
Krambambuli	135	0	507	0.76	4573

Table 1: Corpus statistics and agreement scores across texts in our dataset.

We annotated five German prose texts, all of which were initially published between 1797 and 1915. Following previous work in computational literary studies [21], our annotation units are individual verbal phrases. We instructed three annotators, university students of literary studies, to annotate verb classes in verb phrases. Each verb class is determined by the main verb; modal verbs are ignored. Annotators were encouraged to consult the GermaNet sense inventory via a web interface.[3] They received no explicit class definitions beyond the explanation on the GermaNet website.

Our process involves annotating individual verbs with their corresponding verb class, utilizing GermaNet as a resource during this process. In cases where GermaNet does not cover a sense, the annotation task can become challenging. For example, the sense of the lemma "legen" (to lay) in the sentence "Das Huhn legt ein Ei." (The chicken lays an egg) is not covered by GermaNet, despite covering eight specific senses and six verb classes. In these cases, annotators must select the correct class unaided.

Each text was annotated in full by at least one annotator; to collect agreement scores, additional passages of each text were annotated by a second and, in three out of five cases, a third annotator. In Table 1, we list statistics on the annotations.

[2] https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/ressources/lexica/germanet/description/verbs/verb-semantic-fields/

[3] https://weblicht.sfs.uni-tuebingen.de/rover/

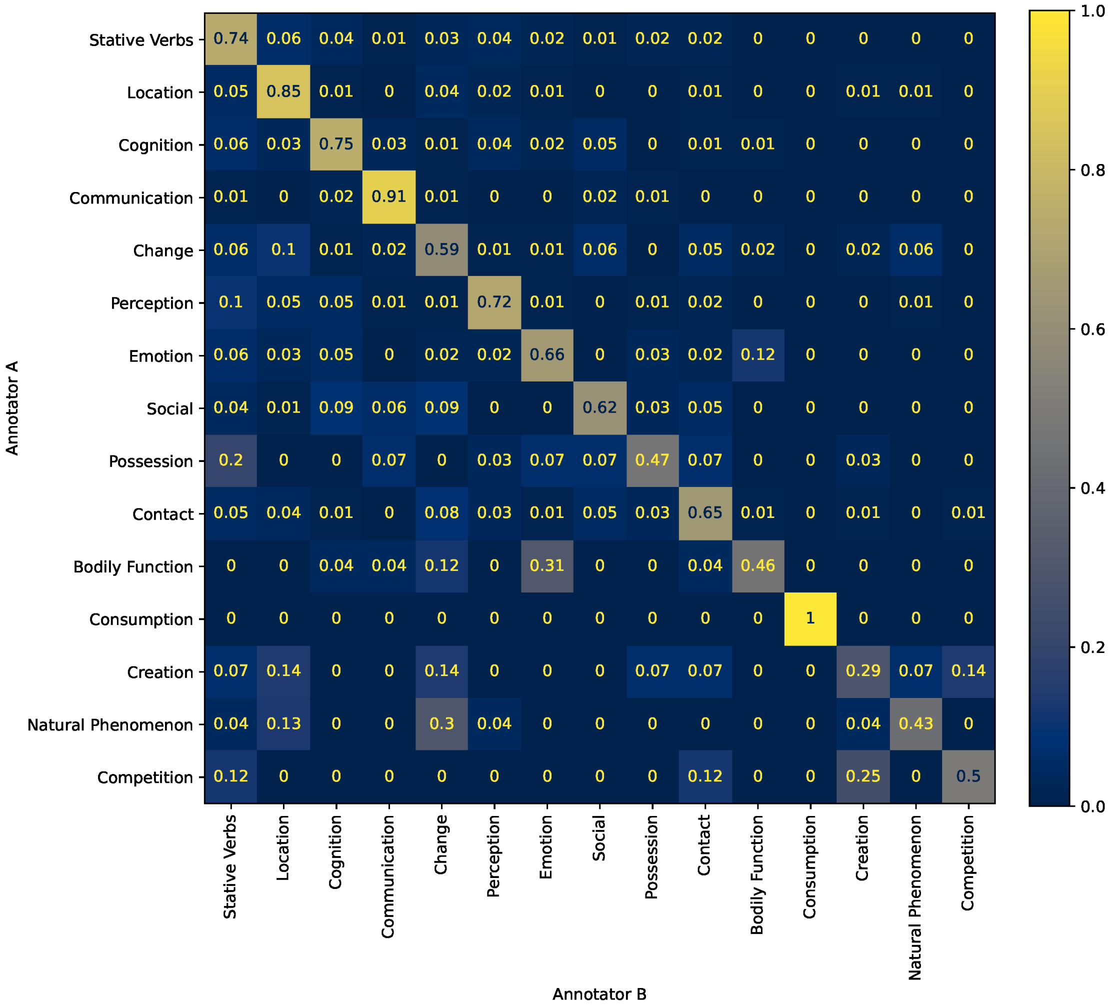

Figure 1: The confusion matrix for all spans with multiple annotations shows annotator (dis)agreement across all texts and annotator pairs. The values are row-normalized.

3 Data Exploration

Macro-averaging across texts, our annotators achieve a Fleiss' Kappa agreement of 0.73 on our 15-class classification task. As shown in Table 1, we have an agreement 0.64 or above for all texts. This is far from a perfect agreement, but we argue that it is expected in our case, as multiple classes could be considered correct in many cases.

The distribution of verb classes in our dataset is top-heavy, with around a quarter of our 11,545 annotations being of the *Stative Verbs* (Allgemein) category. In contrast, rare categories like *Competition* (Konkurrenz) only represent a small fraction (0.42%) of all annotations. See Appendix A for the detailed breakdown.

Figure 1 shows a clear difference in difficulty across classes for human annotators. While the class *Consumption* (Verbrauch) is identified with a perfect agreement, *Competition* (Konkurrenz) is often also identified as *Communication* (Kommunikation), *Location* (Lokation), *Creation* (Schöpfung), or *Stative Verb* (Allgemein).

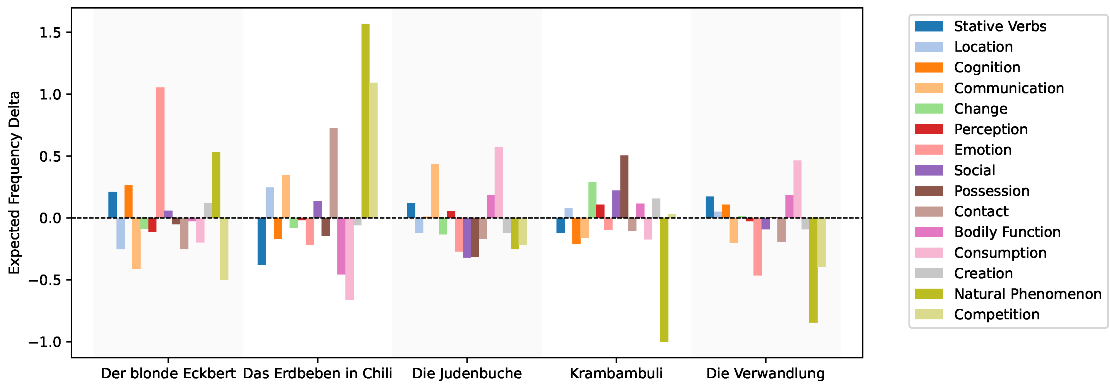

Figure 2: The comparative frequency of verb classes in our analyzed texts, normalized by the expected number of class occurrences given the corpus-wide distribution and the text length.

3.1 Semantic Text Profiles

The diverging bar chart in Figure 2 shows the ratio of verb class occurrences in each of our texts to the expected occurrence of each class given its frequency across all five texts and the given text's length; the plot is based on the data from the main annotator of each text. Since our frequency information is based only on the five texts, meaning the frequencies can be understood only as a comparison with the other texts, a larger reference corpus would be beneficial once automated classification is applied. The text-specific representation of the verb classes reveals discrete semantic profiles in at least four of the five texts. In "Das Erdbeben in Chili", the class of natural phenomena is most prominent. This comes as no surprise, as the title (The Earthquake in Chile) already announces the earthquake that is central in the first part of the text. In contrast, Kafka's "Verwandlung" has the fewest verbs from the class of natural phenomena, which is also not unexpected, as the setting of this story is almost exclusively interior. In the "Krambambuli" text, the dispute between two men over a dog takes the center of attention; here, the highest value is, by a wide margin, found in the *Possession* (Besitz) class. These observations indicate that verb classes can be suitable for the semantic profiling of texts.

In short texts, deviations from the expected value are common in rare classes since just a few instances can lead to the expected value being far surpassed. Aside from the previously mentioned lack of natural phenomena, the plots for the Kafka text and, above all, of "Die Judenbuche" prove to be less distinctive.

Moreover, some common verb classes may be too broad and can contribute to non-distinctive profiles, as evidenced by classes like *Stative Verbs* (Allgemein) and *Cognition* (Kognition), in which we observe little deviation. Finally, we see a sizable spread in the frequency of the *Location* (Lokation) class despite its frequency.

4 Automation

For our initial model comparison, we randomly sample 10 instances from each class, for a total of 150 samples. This stratified sampling approach eliminates bias against underrepresented classes in our evaluation data, thereby improving the robustness of our approach to unseen data with potentially different distributions. While such a bias would improve inference results on our data, it would also impact generalization, especially when estimating the relative class frequencies for entire works on a corpus level. As we consider this sample, our development set, we retain three texts ("Der blonde Eckbert", "Krambambuli", and "Die Verwandlung") for subsequent testing and do not include them in our development set.

Model Name	LLM-arena ELO	Zero-Shot		Few-Shot	
		no RAG	RAG	no RAG	RAG
Random Baseline	-	0.07	-	-	-
Lemma's Random Sense	-	0.41	-	-	-
Llama-3-70B-Instruct [19]	1275	0.43	0.61	0.48	0.68
Gemma-3-27B-it [18]	1365	0.47	0.68	0.49	0.69
GPT-4.1[4]	1412	0.53	**0.75**	0.54	**0.74**

Table 2: Model accuracy in classifying semantic verb classes on our class-balanced development set. For reference, we provide German LM-arena ELO scores as of 2025-07-09.

4.1 Prompting Setups

We test multiple prompting setups to compare their classification performance. All setups are tested in a multi-turn conversion. The first step asks the model to identify the full verb given the manually selected verbal phrase. The second step asks it to reason about the potential verb classes; the third step asks for the actual classification decision. We test adding a retrieval-augmented generation (RAG) [15] component to this setup, which queries GermaNet for the lemma identified in step one. Generally, RAG involves passing a generative model an additional input containing information relevant to the task at hand, often using an embedding-based search. For example, this can mean providing results from a web search to an LLM that is composing the answer to a user's question by including them in the model's input. Our RAG approach searches GermaNet using the lemma previously extracted by the LLM and inputs a textual representation of all synsets for the lemma, including sample usages and definitions. This textual representation is passed to the model alongside the verbal phrase, its lemma, and the context. Thus, in a strict sense, our approach does not conform to the standard RAG pipeline, relying on string search rather than embedding search. We test the effect of in-context few-shot learning for our task.

We focus on testing open-weight models, selecting Llama-3 [19] as a popular choice in the NLP and DH communities and Gemma-3 27B [18], a smaller, more recent model. We test both models at 16-bit precision but acknowledge that quantization could be explored to apply our approach to more data. For comparison, we also test GPT-4, the latest non-reasoning non-preview model by OpenAI.[5]

Table 2 shows the models' performance on our task and their LM arena ELO score [3]. ELO is a win-rate-based measure, in this case based on LLM-arena votes, with users choosing which of two random models' outputs they prefer.

As a baseline, we add an approach that outputs a verb class of random sense out of those present for the lemma. Each lemma typically has a few senses, many of which may share the same verb classes. This baseline already considerably outperforms the random baseline, reaching an accuracy of 0.41 compared to the random baseline's 0.07. As the more recent model, Gemma-3 unsurprisingly exhibits a better arena score than Llama-3 but scores below GPT-4.

The RAG-based approach outperforms the others by far. This raises the question: does the model only disambiguate between existing word senses, or does it sometimes use semantic classes different from those present in GermaNet? To this end, we evaluate the accuracy of our model given specific coverage scenarios. The results are based on lemmas that are automatically extracted by the LLM. Either there are definitions for a given lemma ("Lemma and Class Covered"), or there are no matching definitions ("Lemma not Covered"). In the former case, there is also the option that none of the definitions in GermaNet have the correct verb class ("Lemma Covered but not

[5] Specifically, we use gpt-4.1-2025-04-14.

	Accuracy		
	RAG + Few-Shot	RAG	Few-Shot
Lemma and Class Covered	79.77%	76.40%	49.43%
Lemma not Covered	50.00%	50.00%	50.00%
Lemma Covered but not Class	54.23%	59.32%	47.46%

Table 3: We list the accuracy of Gemma-3 given specific GermaNet coverage scenarios. Of our 150 development set examples, only two lemmas are not covered at all in GermaNet.

Class").

Table 3 shows these results for the Gemma-3 model. We can observe that, given an example with the correct class for a given lemma is present in GermaNet, the chance of a correct output is almost 80%. If no definitions are available for the lemma, the performance drops to around 70%. In the worst case, however, senses are available for a lemma, but none offer the correct verb class, meaning the model has to select a class not covered by the definitions. In our setup, Gemma-3 still achieves a 54.23% accuracy for those examples. We observe that the RAG approach with few-shot examples yields increased accuracy compared to the RAG approach only when the lemma and class are covered. Given that the RAG few-shot examples specifically instruct the model to check all senses individually, this is hardly surprising. In conclusion, we observe that the model disambiguates between existing senses and even picks the correct senses when the retrieval results are misleading in that they only contain other senses.

In preliminary experiments, the reasoning models o3[6] and DeepSeek-R1 [5] did not achieve competitive performance; we suspect that this may be caused by our prompting setup. For our further experiments, we select Gemma-3 with RAG and few-shot examples, although there is no appreciable difference to the setup without few-shot data. While GPT-4.1 performs better than the open-weight models we test, it is cost-prohibitive for our application.[7]

4.2 Results

In this section, after exploring the capability of multiple models on a balanced dataset, we present the result of our automation approach on a held-out test set with natural class distribution. Here, we only consider spans with two or more manual annotations. Table 4 shows the accuracy of all annotators measured by comparing their labels to the predictions of all human annotators (micro-averaged). This effectively serves as an inter-annotator agreement: annotators with high accuracy agree with many annotations. Due to limited annotation resources and a requirement for a sizable dataset to adequately cover rare classes, no gold standard was created. Accordingly, the only way to evaluate the performance of a given annotator is to assess how many of the others agree with them. This also means that perfect accuracy is not attainable for anyone, given that two other annotators disagree on one or more classification across the dataset. To more accurately measure the exact performance of our automated approach, we would have to create gold annotations.

We have not performed extensive out-of-domain validation, so the automated approach could conceivably fail on some other data. However, it is worth noting that our task-specific training data is limited to the few-shot examples we provide.

[6] `https://platform.openai.com/docs/models/o3`
[7] We extrapolated an expected cost of around $150 for inference on the corpus introduced in section 5.

Data Split	Annotator	Accuracy
Test	Annotator 1	0.81
	Annotator 2	0.81
	Annotator 3	0.76
	Gemma 3	0.65

Table 4: Accuracy score for all pairs of predictions in our test set when considering any matching annotation as a gold label. Note that the test set is not class-balanced.

5 Corpus Application

To explore how the distribution of verb classes in texts may indicate genre patterns, we extracted four subcorpora from the d-Prose 1870-1920 corpus [8]. This corpus comprises 2,511 German-language prose texts published between 1870 and 1920, including both popular and highbrow literature. Our subcorpora comprise 29 texts, totaling approximately 1.3 million tokens. In terms of individual length, the texts range from just over 1,000 tokens to around 144,000 tokens. For a full list of all works, see Appendix D. We formed the four genre subcorpora (adventure, crime,

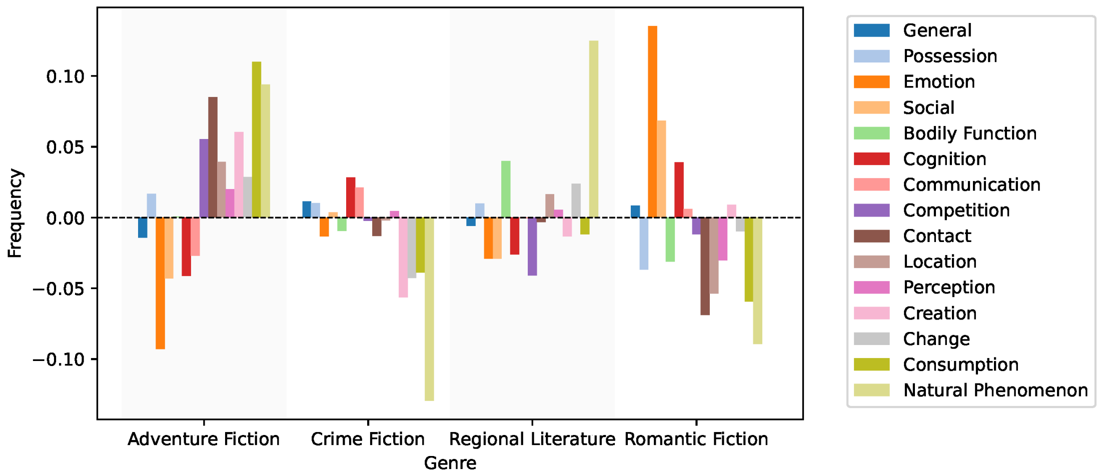

Figure 3: Normalized frequency of each verb class in the four subcorpora. The expected frequency for each class is based on the corpus-wide frequency of the class.

regional literature, and romantic fiction) based on attributions in existing research, selecting the genres according to their popularity during this period. Applying our automated method (section 2) yielded the class distributions shown in Figure 3; verb class frequencies are normalized to account for subcorpus sizes and the corpus wide frequency of each class. The bar chart exhibits some genre-specific outcomes: In romantic fiction, there is a markedly higher prevalence of verbs related to emotion, society, and cognition, reflecting the genre's focus on feelings and interpersonal relationships. The notably lower proportions of contact and bodily function classes may stem from prevailing social conventions and taboos of the era, which necessitated avoiding overt depictions of physical intimacy or sexuality. Thus, romantic fiction of the time tended to feature idealized, emotionally and morally charged – but not erotic – portrayals of love. Romantic and crime fiction both show elevated frequencies of cognition verbs, reflecting a shared emphasis on psychological processes. In crime fiction, the prevalence of cognitive verbs highlights mental activities like thinking, suspicion, inference, and character profiling. The high frequency of communicative verbs corresponds to genre-typical elements such as interrogations, witness statements, or dialogues among investigators [17]. In our crime fiction corpus, natural phenomena play a below-average role – perhaps due to the urban, social, or institutional contexts of most plots – whereas verbs denoting natural phenomena dominate in both adventure fiction and regional literature. This difference in distribution aligns with genre expectations: regional literature often glorifies or provides identity-forming descriptions of the native environment, and its emphasis on social integration and interpersonal relations [2] correlates with frequent use of verbs from the society and contact classes. Our analysis characterizes adventure fiction by high proportions of verbs in the *Location*, *Change*, *Bodily Function*, *Consumption*, and *Natural Phenomenon* classes. These reflect core narrative features of the genre: movement, physical challenges, and encounters with extreme, often nature-driven situations [13]. The prevalence of location verbs may point to frequent changes of place, travel, and exploration of unknown spaces – central motifs of adventure narratives – while abundant change verbs signal the dynamic character of texts in which characters, situations, or external circumstances continually transform.

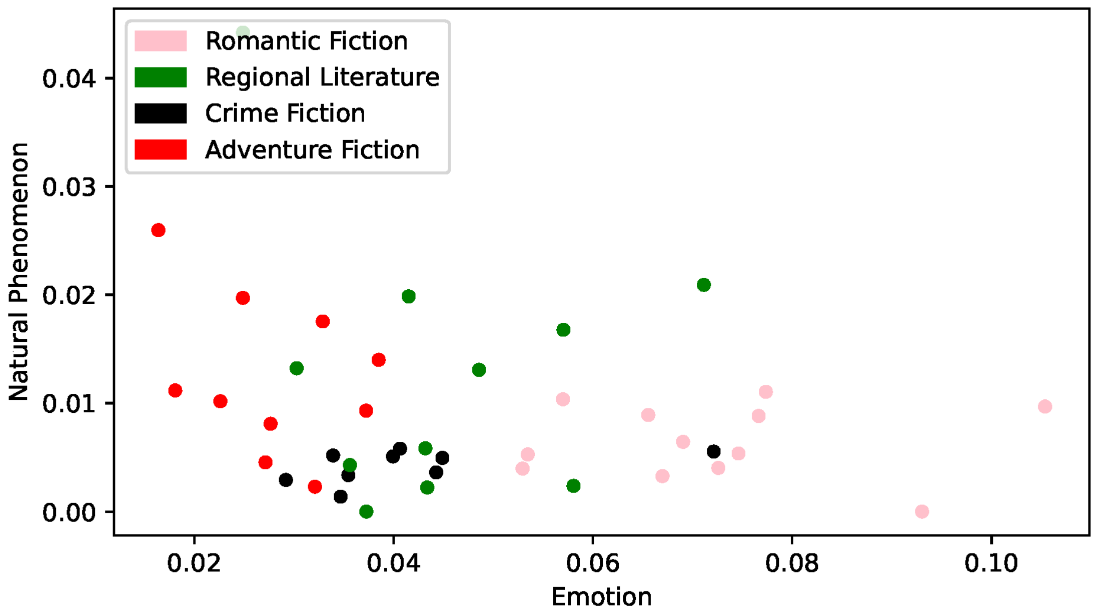

Figure 4: The relative frequencies of select verb classes (here *Natural Phenomenon* and *Emotion*) can exhibit clear differences in their distribution across genres.

Figure 4 illustrates that the frequency profile of verb classes can surface genre-defining features. The plot visualizes an opposition between romantic and adventure fiction (emotional vs. nature-oriented). Crime fiction and regional literature occupy intermediate positions, with regional literature positioned closer to adventure narratives, due to the latter's emphasis on natural phenomena. The clear separation of love stories and adventure stories suggests that verb class profiles can serve as a defining feature in genre classification. A more robust analysis, based on a larger corpus, would be required to definitively indicate that verb classes are a sufficient feature for genre classification.

6 Conclusion

In this work, we showcase an approach to the scalable semantic profiling of texts. We employ LLMs to resolve the classes of verbs in context and build class frequency profiles on a text or corpus level. The LLM inference in our approach is computationally expensive and can be costly to apply to large text corpora, but can yield robust representations of a text's themes. For the LLM evaluation, it is essential to emphasize that the performance of all models is potentially sensitive to the choice of prompt and the exact task formulation; no conclusions as to each model's generalized performance characteristics can be made.

Our exploratory data analysis shows that our approach can detect not only genre-typical actions and themes within texts and corpora, but may also support genre comparison and the diachronic investigation of genre or genre traditions, as demonstrated by the absence of body-related verbs in romantic fiction from the d-Prose corpus and thus from the late 19th and early 20th centuries.

For extended usage of our approach, we intend to explore two avenues of reducing the computational effort: (a) testing quantized versions of the current models and (b) training smaller models in a student-teacher setup to perform the task more efficiently at inference time.

In the future, we aim to apply this method to larger corpora, analyzing additional phenomena, such as literary movements and diachronic analysis. We believe our approach can serve as an interpretable yet reliable exploration tool for literary scholars, while also providing a feature set on which to build text classification approaches for phenomena like genre or literary movement.

References

[1] Baker, Collin F., Fillmore, Charles J., and Lowe, John B. "The Berkeley FrameNet Project". In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. 1998, pp. 86–90. URL: `https://aclanthology.org/C98-1013` (visited on 07/02/2024).

[2] Braungart, Georg, Fricke, Harald, Grubmüller, Klaus, Müller, Jan-Dirk, Vollhardt, Friedrich, and Weimar, Klaus. "Heimatliteratur". In: *Reallexikon der deutschen Literaturwissenschaft*, ed. by Georg Braungart, Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, Friedrich Vollhardt, and Klaus Weimar. Vol. 1. Artikel im Band 1 (A – G). Berlin / New York: Walter de Gruyter, 2010, pp. 19–20. ISBN: 978-3-11-010896-5. DOI: `10.1515/9783110914672`.

[3] Chiang, Wei-Lin et al. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference". Mar. 2024. DOI: `10.48550/arXiv.2403.04132`. arXiv: `2403.04132 [cs]`. (Visited on 06/20/2024).

[4] Daemmrich, Horst S. and Daemmrich, Ingrid G. *Themen Und Motive in Der Literatur. Ein Handbuch*. Tübingen, Basel: Francke, 1995.

[5] DeepSeek-AI et al. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". 2025. arXiv: `2501.12948 [cs.CL]`. URL: `https://arxiv.org/abs/2501.12948`.

[6] Christiane Fellbaum, edited by. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge, MA: MIT Press, 1998. ISBN: 978-0-262-06197-1.

[7] Frenzel, Elisabeth. *Motive Der Weltliteratur: Ein Lexikon Dichtungsgeschichtlicher Längsschnitte*. Stuttgart: Kröner, 1980.

[8] Gius, Evelyn, Guhr, Svenja, and Adelmann, Benedikt. *d-Prose 1870-1920 (2.0) [Data set]*. Zenodo, 2021. DOI: `10.5281/zenodo.5015008`. URL: `https://doi.org/10.5281/zenodo.5015008`.

[9] Hamp, Birgit and Feldweg, Helmut. “GermaNet - a Lexical-Semantic Net for German”. In: *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, ed. by Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorik Wilks. Madrid, Spain, 1997, pp. 9–15. URL: https://aclanthology.org/W97-0802 (visited on 06/20/2024).

[10] Hatzel, Hans Ole, Stiemer, Haimo, Biemann, Chris, and Gius, Evelyn. “Machine Learning in Computational Literary Studies”. In: *it - Information Technology 65*, no. 4-5 (Aug. 2023), ed. by Hubert Mara and Ricardo Usbeck. ISSN: 2196-7032. DOI: 10.1515/itit-2023-0041. (Visited on 08/25/2023).

[11] Hatzel, Hans Ole, Stiemer, Haimo, Biemann, Chris, and Gius, Evelyn. “Towards a Verb Class-based Semantic Analysis of German Literary Texts”. In: *DH2025 Book of Abstracts*. Accepted. Lisbon, Portugal, 2025.

[12] Klavans, Judith L. “The Role of Verbs in Document Analysis”. arXiv:cmplg/9807002. Montreal, Canada, Aug. 1998. DOI: 10.48550/arXiv.cmp-lg/9807002.

[13] Martin von Koppenfels and Manuel Mühlbacher, edited by. *Abenteuer: Erzählmuster, Formprinzip, Genre*. Vol. 1. Philologie des Abenteuers. Paderborn: Wilhelm Fink Verlag, 2019. ISBN: 978-3-7705-6472-9. DOI: 10.30965/9783846764725.

[14] Levin, Beth. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press, Sept. 1993. ISBN: 978-0-226-47533-2. URL: https://press.uchicago.edu/ucp/books/book/chicago/E/bo3684144.html (visited on 06/20/2024).

[15] Li, Huayang, Su, Yixuan, Cai, Deng, Wang, Yan, and Liu, Lemao. “A Survey on Retrieval-Augmented Text Generation”. Feb. 2022. DOI: 10.48550/arXiv.2202.01110. arXiv: 2202.01110 [cs]. (Visited on 07/01/2024).

[16] Maienborn, Claudia. “Event semantics”, ed. by Claudia Maienborn, Klaus von Heusinger, and Paul Portner. Berlin / New York, 2011. DOI: 10.1515/9783110589245-008.

[17] Parra Membrives, Eva and Brylla, Wolfgang Damian. *Facetten des Kriminalromans: Ein Genre zwischen Tradition und Innovation*. 1st ed. Tübingen: Narr Francke Attempto, 2015. ISBN: 978-3-8233-6946-2.

[18] Team, Gemma. “Gemma 3 Technical Report”. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.

[19] Team, Llama-3. “The Llama 3 Herd of Models”. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[20] Uglanova, Inna and Gius, Evelyn. “The Order of Things. A Study on Topic Modelling of Literary Texts”. en. In: *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, ed. by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers. CEUR Workshop Proceedings. Amsterdam, the Netherlands, Nov. 2020, pp. 57–76. URL: http://ceur-ws.org/Vol-2723/long7.pdf.

[21] Vauth, Michael and Gius, Evelyn. “Richtlinien Für Die Annotation Narratologischer Ereigniskonzepte”. July 2021. DOI: 10.5281/zenodo.5078174. (Visited on 07/08/2021).

A Dataset Statistics

Category	Translation	Count	Percentage of Instances
Allgemein	Stative Verbs	2811	24.35%
Lokation	Location	2088	18.09%
Kognition	Cognition	1263	10.94%
Kommunikation	Communication	1108	9.60%
Veränderung	Change	1009	8.74%
Perzeption	Perception	803	6.96%
Gefühl	Emotion	559	4.84%
Gesellschaft	Social	555	4.81%
Besitz	Possession	392	3.40%
Kontakt	Contact	357	3.09%
Körperfunktion	Bodily Function	291	2.52%
Schöpfung	Creation	96	0.83%
Verbrauch	Consumption	83	0.72%
Naturphänomen	Natural Phenomenon	82	0.71%
Konkurrenz	Competition	48	0.42%

Table 5: Verb classes ordered by number of occurrences across all our annotations.

B Prompts

Our prompting setup is performed in multiple steps. In all prompts, we show the "..." notation indicates a response being generated by the LLM, whereas all other messages are passed in as input.
By using the Ollama chat API we apply the appropriate chat template for each model. Other than that, we use no model-specific prompts. The few-shot examples, provided only in the few-shot setup, are customized in two variants depending on whether RAG is enabled. The RAG few-shot examples directly reference the GermaNet senses and provide a textual description of checking every sense.

B.1 Lemma Extraction

The first, lemma extraction is always performed in a few-shot setup using the following prompt, with `<Phrase>` replaced by the phrase in question. For the lemma extraction, after the system prompt we provide a set of 19 few-shot examples, which we omit here for brevity.

System: Du bist ein hilfreicher Assistent.

```
<Few-Shot Examples>
```

User: Was ist das syntaktisch obserste Vollverb in der Phrase '`<Phrase>`'? Bitte antworte nur mit dem Lemma des Vollverbs

---

Assistant: ...

B.2 Classification

For the verb class identification itself we use the following prompt, where `<List Of Classes>` is a list of the names of all classes in German. Additionally, we now include a few sentences of context (`<Context>`) to enable a better disambiguation. We found that instructions in the system prompt were better at constraining the model to this output, additionally we used minimal edit-distance to the gold classes to be able to assign any output to a valid class.

> **System:** Du bist ein Annotator mit linguistischem Expertenwissen und annotierst Verbklassen nach dem Schema von GermaNet. Auf eine Frage die darum bittet nur mit der Klasse zu antworten, antworte stehts nur mit einer der folgenden Klassen: `<List Of Classes>`
>
> ---
>
> `<Few-Shot Examples>`
>
> ---
>
> **User:** Gegeben folgenden Kontext: '`<Context>`' Um welche Klassen könnte es sich bei dem Lemma `<Lemma>` in der Phrase: '`<Phrase>`' handeln? Erläutere deine Entscheidung.
>
> ---
>
> **Assistant:** ...
>
> ---
>
> **User:** Alles in allem, welche der 15 Klassen ist es also wohl? Antworte nur mit der Klasse!
>
> ---
>
> **Assistant:** ...

C Few-Shot Examples

As a sample, we provide one example for each setup (in-context few-shot learning with and without RAG). The actual model is always provided five examples, one each for the classes *Lokation*, *Allgemein*, *Perception*, *Kognition*, and *Kontakt*. The system prompt is omitted for brevity.

In the RAG setup our example explicitly checks all provided senses against the example in question.

> **User:** Gegeben folgenden Kontext: [...] Er senkte sich so tief, daß seine Stirn den Boden berührte, Gott für seine wunderbare Errettung zu danken; [...]
> Um welche Klassen könnte es sich bei dem Lemma 'berühren' in der Phrase: 'daß seine Stirn den Boden berührte' handeln? Erläutere deine Entscheidung.
> Zur Referenz, GermaNet enthält 3 Einträge zum Lemma 'berühren':
>
> 1. 'berühren' aus der Verbklasse Kontakt
> Definition: jemanden / etwas berühren; mit dem Äußeren seines Körpers absichtlich mit etwas Fassbarem in Berührung kommen
> Eine Besipielhafte verwendung wäre: 'Sie berührte ihn am Arm.'
>
> 2. 'berühren' aus der Verbklasse Gefuehl
> Definition: jemanden berühren; zur Trauer oder Nachdenklichkeit anregen Eine Besipielhafte verwendung wäre: 'Das Theaterstück hat mich berührt.' Synonyme sind: anrühren, ergreifen, beseelen, rühren
>
> 3. 'berühren' aus der Verbklasse Kommunikation
> Definition: –
> Eine Besipielhafte verwendung wäre: 'Sie berührten das Phänomen bei der Besprechung nicht.'
>
> ---
>
> **Assistant:** Das Lemma 'berühren' hat drei Einträge in GermaNet. Ich prüfe für alle, ob sie zu unserem Beispiel passen.
> 1. Ja: Die Definition erscheint passend, es geht bei uns um das Berühren der Stirn mit dem Boden. Also mit dem Äußeren des Körpers etwas "Fassbares" zu berühren.

2. Nein: Es geht in unserem Beispiel nicht um eine emotionale Berührtheit wie in diesem Eintrag.
3. Nein: Bei unserem Beispiel geht es nicht um Berühren im Sinne dessen, ein Thema anzusprechen.
Nur 1. erscheint passend, und die dazugehörige Verbklasse ”Kontakt” passt auch eindeutig zu unserem Beispiel.

User: Alles in allem, welche der 15 Klassen ist es also wohl? Antworte nur mit der Klasse!

Assistant: Kontakt

In the few-shot setup without RAG the few-shot example instead argues based on the class semantics without additonal guidance:

User: Gegeben folgenden Kontext: [...] Er senkte sich so tief, daß seine Stirn den Boden berührte, Gott für seine wunderbare Errettung zu danken; [...] Um welche Klassen könnte es sich bei dem Lemma ’berühren’ in der Phrase: ’daß seine Stirn den Boden berührte’ handeln? Erläutere deine Entscheidung.

Assistant: Die Berührung bezieht sich hier auf das physische berühren des Bodens mit der Strin. Derartiger physischer Kontakt ist mit der Klasse ’Kontakt’ zu annotieren.

User: Alles in allem, welche der 15 Klassen ist es also wohl? Antworte nur mit der Klasse!

Assistant: Kontakt

D Genre Corpus

Genre	Author	Title [English Translation]	Year of Publication
Adventure fiction	Ernst Constantin	Das warme Polarland [The Warm Polar Land]	1895
	Friedrich Gerstäcker	Das Wrack [The Wreck]	1864
	Johannes Gillhoff	Juernjakob Swehn, der Amerikafahrer [Juernjakob Swehn, the American Traveler]	1917
	Robert Kraft	Der Medizinmann [The Medicine Man]	1900
	Karl May	Der Oelprinz [The Oil Prince]	1893
	Karl May	Eine Seehundsjagd [A Seal Hunt]	1891
	Karl May	Im wilden Westen Nordamerikas [In the Wild West of North America]	1890
	Friedrich Joachim Pajeken	Bob, der Fallensteller [Bob the Trapper]	1899
	Alois Theodor Sonnleitner	Die Höhlenkinder im heimlichen Grund [The Cave Children in the Hidden Ground]	1907
	Fedor von Zobelitz	Der Kurier des Kaisers [The Emperor’s Courier]	1908
Crime fiction	Matthias Blank	Der Mord im Ballsaal [The Murder in the Ballroom]	1909
	Adalbert Goldscheider	Detektiv Dagobert. Eine Verhaftung [Detective Dagobert: An Arrest]	1910
	Auguste Groner	Der rote Merkur [The Red Mercury]	1910
	Jenny Hirsch	Ein seltsamer Fall [A Strange Case]	1876
	Alfred Schirokauer	Die graue Macht [The Grey Power]	1910
Regional literature	Ludwig Ganghofer	Der Jaeger von Fall [The Hunter of Fall]	1883
	Ludwig Ganghofer	Der Weissbacher [The Weissbacher]	1906
	Wilhelm Jensen	Vor der Elbmündung [At the Elbe Estuary]	1905
	Timm Kröger	Auf der Heide [On the Heath]	1891
	Hermann Löns	Die Häuser von Ohlenhof [The Houses of Ohlenhof]	1917
	Peter Rosegger	Der Pfarrersbub [The Parson’s Boy]	1880
Romantic fiction	Elisabeth Bürstenbinder	Adlerflug [Eagle’s Flight]	1886
	Hermann Heiberg	Küsse [Kisses]	1896
	Hermann Sudermann	Jolanthes Hochzeit [Jolanthe’s Wedding]	1892
	Arthur Zapp	Zapp Arthur Das Liebesleben [The Love Life]	1920
	Stefan Zweig	Die Liebe der Erika [Erika’s Love]	1904
	Karl Emmerich Robert von Bayer	Lydia [Lydia]	1879
	Eduard Graf von Keyserling	Bunte Herzen [Colorful Hearts]	1909
	Ferdinand Ludwig Adam von Saar	Requiem der Liebe [Requiem of Love]	1908

Table 6: Our genre corpus consists of four genres with all works being drawn from the d-Prose 1870-1920 corpus [8].