Artificial Relationships in Fiction: A Dataset for Advancing NLP in Literary Domains

Despina Christou
School of Informatics,
Aristotle University of Thessaloniki,
54124, Greece
christoud@csd.auth.gr

Grigorios Tsoumakas
School of Informatics,
Aristotle University of Thessaloniki,
54124, Greece,
Archimedes, Athena Research Center, Greece,
greg@csd.auth.gr

Abstract

Relation extraction (RE) in fiction presents unique NLP challenges due to implicit, narrative-driven relationships. Unlike factual texts, fiction weaves complex connections, yet existing RE datasets focus on non-fiction. To address this, we introduce *Artificial Relationships in Fiction* (ARF), a synthetically annotated dataset for literary RE. Built from diverse Project Gutenberg fiction, ARF considers author demographics, publication periods, and themes. We curated an ontology for fiction-specific entities and relations, and using GPT-4o, generated artificial relationships to capture narrative complexity. Our analysis demonstrates its value for finetuning RE models and advancing computational literary studies. By bridging a critical RE gap, ARF enables deeper exploration of fictional relationships, enriching NLP research at the intersection of storytelling and AI-driven literary analysis.

1 Introduction

Relation extraction is a fundamental NLP task that identifies and categorizes semantic relationships between entities in text (Wadhwa et al., 2023). While RE has been extensively studied in structured domains like news articles and scientific literature (Zhao et al., 2024), its application to fiction remains underexplored (Bamman et al., 2019). Fictional narratives present unique challenges due to their narrative-driven structures, implicit relationships, and varied linguistic styles that differ significantly from factual texts (Elsner, 2012).

To address this gap, we introduce *Artificial Relationships in Fiction (ARF)*, a synthetically annotated dataset for RE in literary texts. The dataset is constructed from a curated selection of real literary texts sourced from Project Gutenberg, with relationship annotations generated using GPT-4o. Unlike traditional datasets that rely on manual annotation, ARF leverages AI-assisted annotation to extract meaningful relationships within fictional narratives. This approach enables large-scale dataset creation while capturing the complexities of fictional interactions (Yang et al., 2024; Chen et al., 2019).

Our contributions include: (1) introducing a synthetically annotated dataset for RE in fiction, (2) developing a systematic dataset creation methodology combining curated selection with language model-assisted relationship generation, and (3) providing a thorough dataset analysis along with example use cases demonstrating its potential for RE research in fiction.

The paper is structured as follows: Section 2 reviews related work, Section 3 outlines dataset creation, Section 4 presents dataset analysis, Section 5 explores evaluation methods and applications, and Section 6 concludes with future research directions.

2 Related Work

Research on RE has traditionally focused on structured, factual texts. Numerous datasets and approaches address newswire (Zeng et al., 2014; Zhang et al., 2017), biomedical (Gu et al., 2016), finance (Vela and Declerck, 2009), legal (Andrew, 2018), and scientific literature (Luan et al., 2018). For instance, the TACRED dataset (Zhang et al., 2017) provides a large-scale corpus of annotated sentences for relation classification, while the ACE dataset (Doddington et al., 2004) supports multilingual entity, relation, and event detection. Shared tasks at SemEval (Hendrickx et al., 2010) further drive benchmarks, enabling methods from feature-based learning (Mintz et al., 2009) to deep neural networks (Zeng et al., 2014) to advance RE in non-fiction.

Despite these developments, fiction remains comparatively underexplored (Moretti, 2011; Zhang, 2024). Unlike factual prose, fictional narratives often convey relationships implicitly through figurative language, complex story arcs, and evolv-

Proceedings of LaTeCH-CLfL 2025, pages 130–147

ing character dynamics. Early computational literary analysis studied character networks and narrative structures (Moretti, 2011), while systems like BookNLP (Bamman et al., 2014) aided entity extraction and coreference resolution. LitBank (Bamman et al., 2019) annotates literary entities but lacks focus on fictional relationships. Other fiction-oriented datasets target social networks (Hamilton et al., 2025) or characterization (Soni et al., 2023; Bamman et al., 2014) but do not systematically address RE. Limited RE efforts in fiction, e.g., character-location associations (He et al., 2013; Vala et al., 2015; Iyyer et al., 2016; Srivastava et al., 2016; Chaturvedi et al., 2017; Mani et al., 2008), often have narrow scopes and lack comprehensive ontologies that capture the diverse range of fictional entities and relations (Christou and Tsoumakas, 2021; Soni et al., 2023).

In response to this gap, recent work has expanded literary relationship analysis. For instance, (Hamilton et al., 2025) introduced synthetic annotations of social networks in literary texts. However, such efforts typically focus on specific relationship types or small corpora. Advancing RE in fiction requires richer ontologies encompassing characters, settings, objects, abstract concepts, and thematic linkages (Bamman et al., 2019).

Meanwhile, the use of large language models for generating synthetic training data has gained momentum as a means to overcome the scarcity and cost of human annotations (Wei et al., 2023; Jiang et al., 2024). Xu et al. (2023) show how GPT-3.5 excels at few-shot RE, highlighting the potential of LLMs for creative or domain-specific tasks. Leveraging these models, researchers can create datasets that reflect the complexity of fictional narratives while still maintaining consistency and diversity in annotations.

Against this backdrop, we present the Artificial Relationships in Fiction (ARF) dataset, a synthetic resource for literary RE. Using fiction from Project Gutenberg and a tailored ontology, ARF leverages GPT-4o to generate nuanced relationships, enriching resources and advancing research in NLP, storytelling, and computational literary analysis.

3 Dataset Creation

High-quality datasets are vital for NLP, especially in literary domains where relationship extraction requires nuanced understanding but lacks annotations. This section details the creation of *Artificial Relationships in Fiction* through three stages: source selection, chunking, and synthetic relationship generation. The dataset[1] is available in three configurations, each supporting distinct analytical needs in literary NLP.

3.1 Selection Criteria

To ensure broad coverage of fiction subgenres, we curated a diverse set of fiction books from specific Project Gutenberg (PG) bookshelves. The selection process involved:

- **Data Collection**: Extracted all books and their metadata[2] from the following PG bookshelves: Fiction, Children & Young Adult Reading, and Crime/Mystery.
- **Deduplication**: Removed books appearing in multiple bookshelves.
- **Language Filtering**: Retained only English-language books.
- **Copyright Compliance**: Included only books marked as *Public domain in the USA*.
- **Outlier Removal**: Excluded books by authors born before 1300 AD (0.2%) to ensure linguistic consistency. Note that the gap from 1300 to mid 19th c. reflects the absence of fiction books from the specified bookshelves in the source corpus.
- **Text Cleaning**: Fixed encoding mismatches and removed formatting artifacts while preserving paragraph and chapter structure.
- **Metadata Additions**: To support richer fiction analysis, we augmented the dataset with additional metadata:
  Author Gender: Inferred via GPT-4o and manually verified.
  Topic Categorization: Condensed verbose PG subjects[3] into 51 thematic topics for better classification (see Appendix A).

The final dataset, available as `fiction_books` configuration, contains 6,322 unique books written between the mid-19th and mid-20th by 1,716 authors, with a 69%-31% male-female author distribution. Spanning 51 thematic topics, this structured dataset supports literary analysis across genres, authors, and writing styles, facilitating deeper

[1] https://huggingface.co/datasets/Despina/project_gutenberg

[2] PG books extracted metadata: book_id, title, author, author_birth_year, author_death_year, release_date, subjects, language, copyright, text

[3] Example of verbose PG subject: Tarzan (Fictitious character) – Fiction, Africa – Fiction, Fantasy fiction, Good and evil – Juvenile fiction, Adventure stories, Apes – Fiction

insights into thematic relationships and character interactions in fiction.

3.2 Text Chunking

To enable effective relationship extraction, we segmented book texts into five-sentence chunks using a rolling window, where each chunk overlaps by one sentence to maintain coherence. This overlapping strategy helps maintain coherence across segments and ensures that relational mentions extending beyond a single chunk are partially captured. While a five-sentence window limits long-range relationships in literary texts, it balances contextual depth with computational efficiency. The resulting dataset, available as `fiction_books_in_chunks` configuration, comprises 5,961,303 chunks, averaging 943 per book.

3.3 Synthetic Relationship Generation

To improve relationship extraction in fiction, we used GPT-4o to generate synthetic relations for selected PG book chunks within a $1K budget. We subsampled 95,475 chunks while preserving thematic and author-gender distributions (see next chapter for details). Ensuring adherence to a structured ontology was a key priority. Our methodology:

Entity Ontology: Developed the most comprehensive ontology of entity types in literary works to date (see Appendix B).

Relationship Ontology: Designed an ontology capturing nuanced relationships between entity types in fictional narratives (see Appendix C).

LLM-Based Relation Extraction: Constructed a robust GPT-4o prompt (see Appendix D) that integrates entity and relationship ontologies in the system prompt, ensuring relationships are classified strictly within predefined categories. To account for potential deviations, we track inconsistency frequencies and report them in Section 4. Relationships were assumed to exist only between two entities that appear in a span of five sentences. Extracted relationships were formatted as JSON objects to ensure compatibility with computational processing pipelines, including the following fields:

- *entity1*, *entity2*: Related entities' text spans
- *entity1Type*, *entity2Type*: Entities ontology types
- *relation*: Ontology-defined relationship type

Metric	Value
Books Count	96
Authors Count	91
Gender Ratio (M-F)	55%-45%
Subgenres	51
Total chunks	95,475
Avg. Chunks per book	995
Chunks w/o Relations	35,230
Avg. Relations per Book	1337
Avg. Relations per Chunk	1.34

Table 1: Dataset Statistics

4 Dataset Statistics, Evaluation, and Analysis

This section presents an overview of the dataset, including key statistics, examples of extracted relations, and a deeper analysis of its structure. The insights provided here inform potential research directions in NLP applications for fiction.

4.1 Dataset Statistics

The dataset (Table 1) consists of 96 books across 51 fiction subgenres, written by 91 authors with a 55%-45% male-female split, ensuring demographic balance. It includes 95,475 text chunks, averaging 995 per book, with 36.9% containing no explicit relations. On average, each book includes 1,337 relations, while relation-containing chunks feature an average of 1.34 relations, demonstrating a structured relational density. These statistics highlight the dataset's diversity and suitability for NLP tasks such as relation extraction and narrative modeling. A complete list of titles and authors is available in Appendix F.

4.2 Examples of Extracted Relations

To illustrate relation extraction quality, consider the example below, capturing pronoun-based relationships—an established challenge in NLP.

```
[{'entity1': 'Vortigern', 'entity2': 'his
  master's sons', 'entity1Type': 'PER',
  'entity2Type': 'PER','relation': 'enemy_of'},
 {'entity1': 'Vortigern', 'entity2': 'castle',
  'entity1Type': 'PER', 'entity2Type': 'FAC',
  'relation': 'owns'}]
```

Further examples appear in Appendix E. As it can be seen, our curated ontologies and GPT-4o-based prompt extract rich relationships, while smaller models like GPT-4o-mini and spaCy's NER and RE failed to detect these pairs, highlighting our approach's robustness and dataset richness.

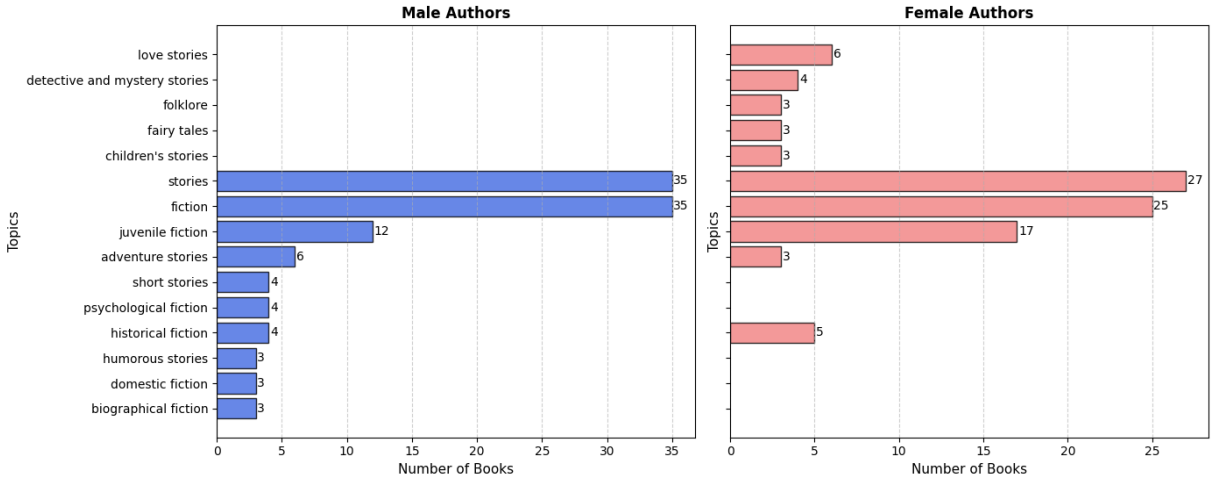

Figure 1: Top-10 subgenres per Author Gender

Ontology Category	New Types	Deviation
Entity 1	10	0.01%
Entity 2	53	0.04%
Relation	3785	2.95%

Table 2: Ontology deviations by category, showing new instances and deviation rates.

4.3 Evaluation

To assess the robustness of the extracted relationships, we analyze deviations from our structured ontology, as summarized in Table 2. This evaluation provides a quantitative measure of the model's accuracy in extracting complex relationships and highlights areas for potential refinement.

Table 2 presents a breakdown of ontology deviations, categorizing inconsistencies observed across entity and relation types. The results indicate that entity deviations are minimal (below 0.05%), while relation-based deviations are comparatively higher (2.95%), likely due to the complexity of multi-faceted relationship extraction. These findings underscore the robustness of our approach while also revealing opportunities for refinement. To further enhance consistency, we plan to incorporate structured output formats from ChatGPT, reducing ambiguity in generated responses.

Additionally, in our preliminary model selection, we tested GPT-4o, GPT-4o-mini, Llama-3.3-70B, Claude-3.5-Sonnet, and Gemini-1.5-Pro for ontology adherence. GPT-4o consistently extracted more complex relationships and was the only model to fully comply without inconsistencies. While this validates our choice, a broader comparative analysis is proposed for future work.

4.4 Analysis

We examine subgenres, entity and relation distributions, and gender-based narrative patterns, providing a foundation for computational literary analysis and NLP in fiction.

4.4.1 Top Subgenres

Figure 1 shows that "fiction," "stories," and "juvenile fiction" dominate across genders. Note, that books often belong to multiple subgenres, reflecting literary fluidity. Despite this shared prominence, distinct gender-based patterns emerge. Male authors favor adventure, humor, and biographical fiction, often engaging with historical and psychological narratives. Female authors emphasize relational and cultural storytelling through love stories, folklore, and children's literature. "Historical fiction" remains a shared interest, suggesting its broad thematic appeal. These patterns provide insights into how gender shapes thematic priorities and narrative structures in fiction in 1850-1950.

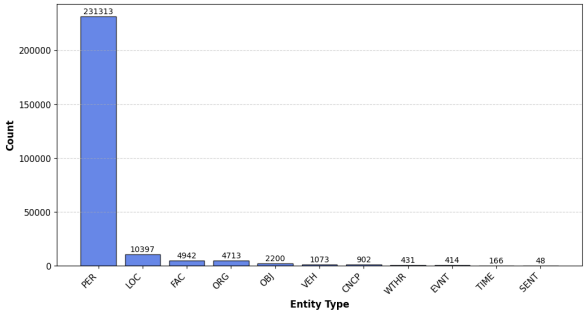

Figure 2: Entity Types Distribution.

4.4.2 Top Entities and Relations

Figures 2, 3 highlight the dataset's character-driven nature. The Person (PER) category dominates with 231,313 occurrences, underscoring the centrality of characters in fiction. Categories such as Location (LOC), Facility (FAC), and Organization (ORG) represent settings and institutions integral to world-building but are far less frequent. Rare entity types like Weather (WTHR), Event (EVNT), and Time (TIME) suggest their secondary importance in storytelling. This distribution highlights the emphasis on characters and their environments in fictional narratives.

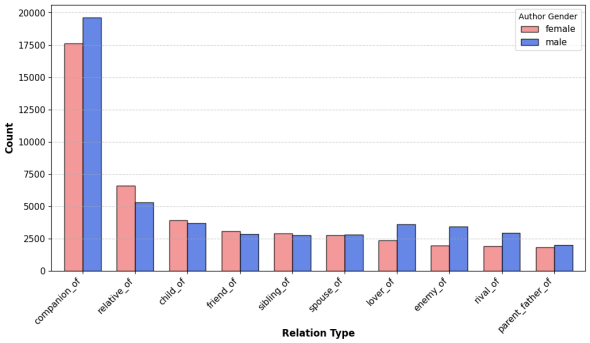

Figure 3: Top-10 relations with author gender usage

Gender-related trends show men slightly favoring PER entities, suggesting a focus on character-driven narratives, while women more frequently use FAC entities, emphasizing settings and contexts. These variations may reflect historical literary norms (Mulvey et al., 2006; Flanagan, 2009).

The most common relations (Figure 3) underscore interpersonal themes. Companion_of is most frequent, highlighting partnership dynamics, alongside familial (relative_of, child_of) and romantic (spouse_of, lover_of) ties. Conflict-driven relations (enemy_of, rival_of) add narrative tension. Gender trends show male-authored works featuring power structures (e.g., kings, warriors), while female-authored works emphasize domestic and relational dynamics. These patterns align with historical literary conventions, shaping how fiction evolved between 1850 and 1950.

5 Use Cases

This dataset offers valuable applications in fiction-specific NLP. It enables model finetuning, helping adapt NLP models for relationship extraction in literary narratives. It also supports literary analysis, allowing researchers to study character networks, relationship evolution, and thematic trends at scale. Additionally, it has creative applications, enhancing AI-driven storytelling and character development for writers, game designers, and digital creators by ensuring richer, more consistent narratives.

6 Limitations & Future Work

While this dataset advances fiction-specific RE, its synthetic nature pose challenges. Below, we outline key limitations and propose future directions to address them.

GPT-4o-generated annotations may introduce biases or inaccuracies, especially for complex or implicit relationships requiring deeper narrative understanding. The reliance on five-sentence chunks, while computationally feasible, limits the capture of long-range relationships across chapters or books, and the absence of explicit coreference resolution hinders tracking evolving character interactions. Without systematic human validation, precision and recall remain unverified, highlighting the need for manual evaluation.

Future work includes a small-scale human validation study, leveraging OpenAI's structured output mode for stricter ontology adherence, and integrating coreference resolution to improve continuity. Adaptive chunking strategies may enhance long-range dependency extraction. Comparative studies with other models and relation extraction systems will assess performance, while active learning could expand the dataset efficiently. Addressing these limitations will enhance reliability and broaden applicability in literary NLP research, enabling deeper narrative analysis.

7 Conclusion

This paper introduces *Artificial Relationships in Fiction*, a synthetically annotated dataset for relation extraction in literary texts. Built from public-domain fiction and GPT-4o generated relationships, ARF bridges structured RE tasks and fictional narratives. Our analysis demonstrates its ability to capture diverse literary relationships, supporting research in character networks, thematic links, and narrative NLP. Challenges include synthetic biases and scalability, requiring future work on human validation and dataset expansion. We envision *ARF* as a foundational resource for NLP, literary analysis, and AI-driven storytelling.

References

Judith Jeyafreeda Andrew. 2018. Automatic extraction of entities and relation from legal documents. In

Proceedings of the Seventh Named Entities Workshop, pages 1–8.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2287–2295.

Despina Christou and Grigorios Tsoumakas. 2021. Extracting semantic relationships in greek literary texts. *Sustainability*, 13(16):9391.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644.

Mary Flanagan. 2009. *Critical play: Radical game design*. MIT press.

J Gu, L Qian, and G Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database: the Journal of Biological Databases and Curation*, 2016:baw042–baw042.

Sil Hamilton, Rebecca MM Hicke, David Mimno, and Matthew Wilkens. 2025. A city of millions: Mapping literary social networks at scale. *arXiv preprint arXiv:2502.19590*.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. GenRES: Rethinking evaluation for generative relation extraction in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2820–2837, Mexico City, Mexico. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. 2008. Spatialml: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Franco Moretti. 2011. Network theory, plot analysis.

Laura Mulvey, Kaja Silverman, Teresa de Laurentis, and Barbara Creed. 2006. Feminist film theorists.

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.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. 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*, pages 769–774.

Mihaela Vela and Thierry Declerck. 2009. Concept and relation extraction in the finance domain. In *Proceedings of the eight international conference on computational semantics*, pages 346–350.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Chatie: Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023. How to unleash the power of large language models for few-shot relation extraction? *arXiv preprint arXiv:2305.01555*.

Yi Yang, Aida Davani, Avirup Sil, and Anoop Kumar. 2024. Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 6: Industry track). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Jiarui Zhang. 2024. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4005–4016, Miami, Florida, USA. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39.

A Thematic Topic Classification

id	Thematic Topic	id	Thematic Topic
1	autobiographical fiction	27	children's stories
2	biographical fiction	28	christmas stories
3	crime fiction	29	code and cipher stories
4	diary fiction	30	college stories
5	didactic fiction	31	cricket stories
6	domestic fiction	32	detective and mystery stories
7	fantasy fiction	33	erotic stories
8	fiction	34	football stories
9	gothic fiction	35	frame-stories
10	historical fiction	36	ghost stories
11	juvenile fiction	37	humorous stories
12	musical fiction	38	hunting stories
13	mystery fiction	39	legal stories
14	paranormal fiction	40	love stories
15	political fiction	41	mystery and detective stories
16	psychological fiction	42	nature stories
17	science fiction	43	opera stories
18	fables	44	railroad stories
19	fairy tales	45	sea stories
20	folklore	46	short stories
21	legends	47	sports stories
22	mythology	48	spy stories
23	tales	49	stories
24	adventure stories	50	war stories
25	baseball stories	51	western stories
26	bible stories		

B Ontology of Entity Types in Fiction

id	Entity Type	Short Description	Description
1	PER	Person	A single person identified by a proper name or a common noun phrase. This category also includes groups or sets of people. Examples: Tom Sawyer, the boy, her daughters, the Ashburnhams.
2	FAC	Facility	A functional, man-made structure created for human use, including spaces for habitation, storage, transportation, and outdoor purposes. Interior spaces like rooms and closets are also included. Examples: the museum, a barn, the highway, the garden, a kitchen.
3	LOC	Location	Physical places without political boundaries, including natural areas, loosely defined regions, or celestial bodies. Examples: the woods, the river, New England, Mars.
4	WTHR	Weather	Natural atmospheric or celestial phenomena, such as storms, droughts, or celestial events. Examples: a thunderstorm, a drought, a solar eclipse, the first snow.

5	VEH	Vehicle	Physical devices designed for transportation, often reflecting historical modes of travel in literature. Examples: a ship, a train, a carriage, a steamboat.
6	ORG	Organization	Formal associations or institutional entities, including administrative, military, political, or religious groups. Examples: the army, the Church (as an organization, not a building), the guild.
7	EVNT	Event	Significant historical, cultural, or personal occurrences within the narrative. Examples: the ball at Netherfield, a proposal in the rain, the war, a festival.
8	TIME	Time Expression	Periods or temporal expressions, including historical eras or chronological markers. Examples: Victorian Era, the Renaissance, the 20th century, a winter evening.
9	OBJ	Object	Artifacts or tangible items of significance within the text. Examples: a letter, a necklace, a sword, a painting.
10	SENT	Sentiment	Emotional states or feelings expressed within the narrative. Examples: happiness, jealousy, anger, grief.
11	CNCP	Concept	Abstract themes or ideas explored in the text, often representing motifs or ideologies. Examples: love, justice, betrayal, courage, freedom.

C Ontology of Relation Types in Fiction

id	Relation Type	Description	Entity1 Type	Entity2 Type
1	parent_father_of	Represents the relationship between a parent and their father. Example: Darth Vader is father_of Luke Skywalker.	PER	PER
2	parent_mother_of	Represents the relationship between a parent and their mother. Example: Cersei Lannister is mother_of Joffrey Baratheon.	PER	PER
3	child_of	Represents the relationship between a child and its parents. Example: Harry Potter is child_of James Potter and Lily Potter.	PER	PER
4	sibling_of	Denotes siblings within the same family. Example: Thor is sibling_of Loki.	PER	PER
5	spouse_of	Indicates a marital relationship, regardless of gender or cultural context. Example: Elizabeth Bennet is spouse_of Mr. Darcy.	PER	PER
6	relative_of	Captures a broader familial connection beyond immediate family, such as cousins, uncles, or distant relatives. Example: Hamlet is relative_of Claudius (uncle-nephew relationship).	PER	PER
7	adopted_by	Indicates a non-biological familial or societal relationship, such as legal guardianship or cultural adoption. Example: Jon Snow is adopted_by Ned Stark.	PER	PER

8	companion_of	A broader term for someone who accompanies, aids, or supports another, including travel companions or loyal allies. Example: Don Quixote is companion_of Sancho Panza.	PER	PER
9	friend_of	Indicates a strong, platonic relationship. Example: Frodo is friend_of Samwise.	PER	PER
10	lover_of	Represents a romantic or amorous relationship, whether mutual or unrequited. Example: Romeo is lover_of Juliet.	PER	PER
11	rival_of	Indicates a competitive relationship that may involve admiration, respect, or antagonism, not necessarily hostile. Example: Sherlock Holmes is rival_of Professor Moriarty.	PER	PER
12	enemy_of	Represents rivalry, hostility, or animosity among people or organizations. Example: Harry Potter is enemy_of Voldemort.	PER/ORG	PER/ORG
13	inspires	To show a motivational or creative influence. Example: Virgil inspires Dante in The Divine Comedy.	PER	PER
14	sacrifices_for	To capture an act of selflessness for another. Example: Sydney Carton sacrifices_for Charles Darnay in *A Tale of Two Cities*.	PER	PER
15	mentor_of	Describes a teaching, guiding, or advisory relationship where one person provides knowledge or support. Example: Dumbledore is mentor_of Harry Potter.	PER	PER
16	teacher_of	To capture formal or academic teaching relationships, distinct from mentor relationships. Example: Snape is teacher_of Harry Potter.	PER	PER
17	protector_of	Represents a caretaking or safeguarding bond, often involving physical or emotional security. Example: Hagrid is protector_of Harry Potter.	PER	PER
18	employer_of	Denotes a work-related hierarchical relationship between an employer and an employee. Example: Ebenezer Scrooge is employer_of Bob Cratchit.	PER	PER
19	leader_of	Indicates a leadership role where an individual leads a group, organization, or nation. Example: Aragorn is leader_of the Fellowship of the Ring.	PER	ORG
20	member_of	Represents membership or affiliation with a group, organization, or society. Example: Harry Potter is member_of Gryffindor House.	PER	ORG
21	lives_in	Specifies a person's residence. Example: Bilbo lives_in Bag End.	PER	FAC/LOC
22	lived_in	Represents historical association. Example: Jane Eyre lived_in the Victorian Era.	PER	TIME
23	visits	Captures temporary presence in a place or facility, such as a visit to a specific location or landmark. Example: Pip visits Satis House in *Great Expectations*.	PER	FAC

24	travel_to	Indicates movement or journey to a specific location, whether planned or incidental. Example: Odysseus travels_to Ithaca.	PER	LOC
25	born_in	A person's birthplace. Example: Napoleon was born_in Corsica.	PER	LOC
26	travels_by	Describes transport modes. Example: Sherlock Holmes travels_by carriage.	PER	VEH
27	participates_in	A person attending or involved in an event. Example: Elizabeth Bennet participates_in the Netherfield Ball.	PER	EVNT
28	causes	A person triggering an event. Example: Macbeth causes Duncan's murder.	PER	EVNT
29	owns	Represents possession of objects. Example: Bilbo owns the Ring.	PER	OBJ
30	believes_in	Represents an individual's ideology, faith, or belief in a concept, philosophy, or ideal. Example: Atticus Finch believes_in justice.	PER	CNCP
31	embodies	A person symbolizing an abstract idea. Example: Beowulf embodies courage.	PER	CNCP
32	located_in	Indicates geographic placement. Example: The Louvre is located_in Paris.	FAC	LOC
33	part_of	Smaller entities within larger ones. Example: The throne room is part_of the castle.	FAC/LOC/ ORG	FAC/LOC/ ORG
34	owned_by	Represents ownership. Example: Thornfield Hall is owned_by Mr. Rochester.	FAC/VEH	PER
35	occupied_by	Indicates current inhabitant. Example: Bag End is occupied_by Frodo.	FAC	PER
36	used_by	Represents organizational usage. Example: The palace is used_by the monarchy.	FAC	ORG
37	affects	Weather affecting a location or an event. Example: The storm affects the village.	WTHR	LOC/ EVNT
38	experienced_by	A person enduring weather. Example: The storm is experienced_by King Lear.	WTHR	PER
39	travels_in	Indicates vehicle operation in specific areas. Example: The ship travels_in the Pacific Ocean.	VEH	LOC
40	based_in	Geographic headquarters. Example: The Knights Templar is based_in Jerusalem.	ORG	LOC
41	attended_by	Persons present at the event. Example: The ball is attended_by Elizabeth Bennet.	EVNT	PER
42	ends_in	To represent temporal conclusions. Example: The war ends_in 1945.	EVNT	TIME
43	occurs_in	The event's geographic location. Example: The battle occurs_in France/spring.	EVNT	LOC/ TIME
44	features	Objects central to the event. Example: The duel features swords.	EVNT	OBJ
45	stored_in	Placement in a specific location. Example: The painting is stored_in the gallery.	OBJ	LOC/FAC
46	expressed_by	Emotional expression. Example: Jealousy is expressed_by Othello.	SENT	PER
47	used_by	Denotes usage. Example: Arthur uses Excalibur.	OBJ	PER

48	associated_with	Concepts tied to events. Example: Justice is associated_with the trial.	CNCP	EVNT

D GPT-4o Prompt for Fictional Relationship Annotation

System Prompt

You are an expert Literature Analyst specializing in identifying entities and their relationships within excerpts from literary works. Your task is to analyze text chunks and extract meaningful **relations** between **entities** based on predefined ontologies below.

Relations Ontology

Each relation connects two entities, defined by their types and descriptions. Use this ontology to categorize relationships accurately.

ID	Relation Type	Description	Entity Type 1	Entity Type 2
1	parent_father_of	Represents the relationship between a parent and their father.	PER	PER
2	parent_mother_of	Represents the relationship between a parent and their mother.	PER	PER
3	child_of	Represents the relationship between a child and their parents.	PER	PER
4	sibling_of	Denotes siblings within the same family.	PER	PER
5	spouse_of	Indicates a marital relationship, regardless of gender or cultural context.	PER	PER
6	relative_of	Captures a broader familial connection beyond immediate family, such as cousins, uncles, or distant relatives.	PER	PER
7	adopted_by	Indicates a non-biological familial or societal relationship, such as legal guardianship or cultural adoption.	PER	PER
8	companion_of	Represents someone who accompanies, aids, or supports another.	PER	PER
9	friend_of	Indicates a strong, platonic relationship.	PER	PER
10	lover_of	Represents a romantic or amorous relationship, whether mutual or unrequited.	PER	PER
11	rival_of	Indicates a competitive relationship that may involve admiration, respect, or antagonism.	PER	PER
12	enemy_of	Represents rivalry, hostility, or animosity among people or organizations.	PER/ORG	PER/ORG

ID	Relation Type	Description	Entity Type 1	Entity Type 2
13	inspires	Shows motivational or creative influence.	PER	PER
14	sacrifices_for	Captures an act of selflessness for another.	PER	PER
15	mentor_of	Describes a teaching, guiding, or advisory relationship.	PER	PER
16	teacher_of	Captures formal or academic teaching relationships.	PER	PER
17	protector_of	Represents a caretaking or safeguarding bond.	PER	PER
18	employer_of	Denotes a work-related hierarchical relationship.	PER	PER
19	leader_of	Indicates a leadership role over a group or organization.	PER	ORG
20	member_of	Represents membership or affiliation with a group or organization.	PER	ORG
21	lives_in	Specifies a person's residence.	PER	FAC/LOC
22	lived_in	Represents historical association.	PER	TIME
23	visits	Captures temporary presence in a place or facility.	PER	FAC
24	travel_to	Indicates movement or journey to a specific location.	PER	LOC
25	born_in	Represents a person's birthplace.	PER	LOC
26	travels_by	Describes transport modes.	PER	VEH
27	participates_in	Captures involvement in an event.	PER	EVNT
28	causes	Represents a person triggering an event.	PER	EVNT
29	owns	Represents possession of objects.	PER	OBJ
30	believes_in	Represents an individual's belief in a concept.	PER	CNCP
31	embodies	Represents a person symbolizing an abstract idea.	PER	CNCP
32	located_in	Indicates geographic placement.	FAC	LOC
33	part_of	Represents smaller entities within larger ones.	FAC/LOC/ORG	FAC/LOC/ORG
34	owned_by	Represents ownership.	FAC/VEH	PER
35	occupied_by	Indicates current inhabitant.	FAC	PER
36	used_by	Represents usage of a facility or object.	FAC	ORG
37	affects	Weather affecting a location or event.	WTHR	LOC/EVNT
38	experienced_by	A person enduring weather.	WTHR	PER

ID	Relation Type	Description	Entity Type 1	Entity Type 2
39	travels_in	Indicates vehicle operation in specific areas.	VEH	LOC
40	based_in	Represents geographic headquarters.	ORG	LOC
41	attended_by	Represents persons present at an event.	EVNT	PER
42	ends_in	Represents temporal conclusions.	EVNT	TIME
43	occurs_in	Represents an event's geographic location or time.	EVNT	LOC/TIME
44	features	Represents objects central to the event.	EVNT	OBJ
45	stored_in	Represents placement of objects in a location.	OBJ	LOC/FAC
46	expressed_by	Represents emotional expression.	SENT	PER
47	used_by	Represents usage of objects.	OBJ	PER
48	associated_with	Represents concepts tied to events.	CNCP	EVNT

Entity Types Ontology

Entities are categorized by their types. Use these definitions to identify and classify entities within the text.

ID	Entity Type	Short Description	Description
1	PER	Person	A single person or group of people.
2	FAC	Facility	Functional, man-made structures for human use.
3	LOC	Location	Physical places without political boundaries.
4	WTHR	Weather	Natural atmospheric or celestial phenomena.
5	VEH	Vehicle	Physical devices for transportation.
6	ORG	Organization	Formal associations or institutions.
7	EVNT	Event	Significant occurrences or actions.
8	TIME	Time Expression	Chronological markers or eras.
9	OBJ	Object	Tangible items of significance.
10	SENT	Sentiment	Emotional states or feelings.
11	CNCP	Concept	Abstract themes or ideas.

Prompt

Identify and extract all related named entity pairs from the provided text. Format the extracted pairs as a list of JSON objects using the structure below for each found relation. Ensure only the list of JSON objects is returned, without any additional text.

```
[
  {
    "entity1": "Exact text of the first entity",
    "entity2": "Exact text of the second entity",
```

```
    "entity1Type": "Type of the first entity",
    "entity2Type": "Type of the second entity",
    "relation": "Relation type"
  }
]
```

Text:

<the text chunk>

E Examples of Extracted Relations

Input Text Chunk	Output Relations
At those words Vortigern's face grew white as ashes, and, rising in confusion and disorder, he sent for all the best artificers and craftsmen and mechanics, and commanded them vehemently to go and build him straightway in the furthest west of his lands a great and strong castle, where he might fly for refuge and escape the vengeance of his master's sons—"and, moreover," cried he, "let the work be done within a hundred days from now, or I will surely spare no life amongst you all." Then all the host of craftsmen, fearing for their lives, found out a proper site whereon to build the tower, and eagerly began to lay in the foundations. But no sooner were the walls raised up above the ground than all their work was overwhelmed and broken down by night invisibly, no man perceiving how, or by whom, or what. And the same thing happening again, and yet again, all the workmen, full of terror, sought out the king, and threw themselves upon their faces before him, beseeching him to interfere and help them or to deliver them from their dreadful work. Filled with mixed rage and fear, the king called for the astrologers and wizards, and took counsel with them what these things might be, and how to overcome them.	{ {'entity1': 'Vortigern', 'entity2': 'his master's sons', 'entity1Type': 'PER', 'entity2Type': 'PER', 'relation': 'enemy_of'}, {'entity1': 'Vortigern', 'entity2': 'castle', 'entity1Type': 'PER', 'entity2Type': 'FAC', 'relation': 'owns'}, {'entity1': 'Vortigern', 'entity2': 'astrologers and wizards', 'entity1Type': 'PER', 'entity2Type': 'PER', 'relation': 'companion_of'} }
"Thou art full young and tender of age," said King Arthur, "to take so high an order upon thee." "Sir," said Griflet, "I beseech thee make me a knight;" and Merlin also advising the king to grant his request, "Well," said Arthur, "be it then so," and knighted him forthwith. Then said he to him, "Since I have granted thee this favour, thou must in turn grant me a gift." "Whatsoever thou wilt, my lord," replied Sir Griflet. "Promise me," said King Arthur, "by the faith of thy body, that when thou hast jousted with this knight at the fountain, thou wilt return to me straightway, unless he slay thee."	{ {'entity1': 'King Arthur', 'entity2': 'Griflet', 'entity1Type': 'PER', 'entity2Type': 'PER', 'relation': 'mentor_of'}, {'entity1': 'Merlin', 'entity2': 'King Arthur', 'entity1Type': 'PER', 'entity2Type': 'PER', 'relation': 'advises'} }

F Dataset Collection - Titles and Authors

PG Book ID	Title	Author

106	Jungle Tales of Tarzan	Edgar Rice Burroughs
12371	The Experiences of a Barrister, and Confessions of an Attorney	Samuel Warren
12753	The Legends of King Arthur and His Knights	James, Sir Knowles
12807	Dick Prescott's Fourth Year at West Point - Or, Ready to Drop the Gray for Shoulder Straps	H. Irving (Harrie Irving) Hancock
1329	A Voyage to Arcturus	David Lindsay
134	Maria; Or, The Wrongs of Woman	Mary Wollstonecraft
14174	The Mating of Lydia	Humphry, Mrs. Ward
15284	The Tale of Johnny Town-Mouse	Beatrix Potter
1574	Historic Girls: Stories Of Girls Who Have Influenced The History Of Their Times	Elbridge S. (Elbridge Streeter) Brooks
1617	The Wind in the Rose-Bush, and Other Stories of the Supernatural	Mary Eleanor Wilkins Freeman
165	McTeague: A Story of San Francisco	Frank Norris
16630	Empire Builders	Francis Lynde
1881	The Call of the Canyon	Zane Grey
18873	Contes et légendes. 1re Partie	H. A. (Hélène Adeline) Guerber
21299	Blue Jackets: The Log of the Teaser	George Manville Fenn
21446	Favourite Fables in Prose and Verse	Harrison Weir
22066	The Long Roll	Mary Johnston
23060	The Unknown Masterpiece - 1845	Honoré de Balzac
24584	Man Overboard!	F. Marion (Francis Marion) Crawford
24714	Fairy Tales from Brazil: How and Why Tales from Brazilian Folk-Lore	Elsie Spicer Eells
25165	The Candy Country	Louisa May Alcott
25205	Light On the Child's Path	William Allen Bixler
25513	Edmund Dulac's Fairy-Book: Fairy Tales of the Allied Nations	Edmund Dulac
2662	Under the Greenwood Tree; Or, The Mellstock Quire - A Rural Painting of the Dutch School	Thomas Hardy
29452	The Wings of the Dove, Volume 1 of 2	Henry James
30365	In Desert and Wilderness	Henryk Sienkiewicz
31217	Household Papers and Stories	Harriet Beecher Stowe
31858	Ancestors: A Novel	Gertrude Franklin Horn Atherton
32543	The White Chief of the Caffres	Alfred W. (Alfred Wilks) Drayson
3322	East Lynne	Henry, Mrs. Wood
33382	Penny Nichols and the Black Imp	Joan Clark

34025	Ancient Rome: The Lives of Great Men	Mary Agnes Hamilton
35179	The Three Sapphires	William Alexander Fraser
35504	Miss Maitland, Private Secretary	Geraldine Bonner
35671	The Messenger	Elizabeth Robins
36684	Molly Brown's Freshman Days	Nell Speed
36703	A Bayard From Bengal - Being some account of the Magnificent and Spanking Career of Chunder Bindabun Bhosh,...	F. Anstey
37121	Charles Dickens' Children Stories	Charles Dickens
37251	In Touch with Nature: Tales and Sketches from the Life	Gordon Stables
39018	Mr. Marx's Secret	E. Phillips (Edward Phillips) Oppenheim
39375	Christmas-Tree Land	Mrs. Molesworth
396	The Lady, or the Tiger?	Frank R. Stockton
40033	The Missing Formula - Madge Sterling Series, 1	Mildred A. (Mildred Augustine) Wirt
40882	Felix Holt, the Radical	George Eliot
42455	The Mystery of the Sea	Bram Stoker
42934	Polly's Southern Cruise	Lillian Elizabeth Roy
43982	Stories of the Old World	Alfred John Church
44	The Song of the Lark	Willa Cather
44111	Red Dynamite - A Mystery Story for Boys	Roy J. (Roy Judson) Snell
4470	Diana of the Crossways — Complete	George Meredith
44872	The Man Who Fell Through the Earth	Carolyn Wells
45517	The Putnam Hall Cadets; or, Good Times in School and Out	Edward Stratemeyer
47139	Stories from Wagner	J. Walker (Joseph Walker) McSpadden
47634	Sons and Lovers	D. H. (David Herbert) Lawrence
5111	The Real Diary of a Real Boy	Henry A. (Henry Augustus) Shute
5182	The Old English Baron: a Gothic Story	Clara Reeve
51919	Rancho Del Muerto, and Other Stories of Adventure - by Various Authors, from "Outing"	Charles King
52610	Ward Hill, the Senior	Everett T. (Everett Titsworth) Tomlinson
52617	The Decameron (Day 1 to Day 5) - Containing an hundred pleasant Novels	Giovanni Boccaccio
52702	Mrs Peixada	Henry Harland
53920	Kittyboy's Christmas	Amy Ella Blanchard

540	The Red Fairy Book	Andrew Lang
55847	Known to the Police	Thomas Holmes
56085	The Silver Princess in Oz	Ruth Plumly Thompson
5658	Lord Jim	Joseph Conrad
56665	Tales and Stories - Now First Collected	Mary Wollstonecraft Shelley
59136	Finkler's Field: A Story of School and Baseball	Ralph Henry Barbour
6053	Evelina, Or, the History of a Young Lady's Entrance into the World	Fanny Burney
61457	Charley's Log: A Story of Schoolboy Life	Emma Leslie
619	The Warden	Anthony Trollope
62126	Captivating Bible Stories for Young People, Written in Simple Language	Charlotte M. (Charlotte Mary) Yonge
64264	Zero Hour	Ray Bradbury
653	The Chimes - A Goblin Story of Some Bells That Rang an Old Year out and a New Year In	Charles Dickens
66687	Fairy Tales for Workers' Children	Hermynia Zur Mühlen
6852	Venus in Furs	Leopold, Ritter von Sacher-Masoch
6941	Old Mortality, Complete	Walter Scott
6985	A Prefect's Uncle	P. G. (Pelham Grenville) Wodehouse
70653	Rattle of Bones	Robert E. (Robert Ervin) Howard
71864	The White Countess	Florence Warden
72063	Once Upon a Time Animal Stories	Carolyn Sherwin Bailey
72824	The Mystery of the Blue Train	Agatha Christie
73548	The Story of the Rhinegold (Der Ring des Nibelungen) Told for Young People	Anna Alice Chapin
74155	A Frontier Knight: A Story of Early Texan Border-Life	Amy Ella Blanchard
74440	Two Brave Boys, and, The Wrong Twin	Mary E. (Mary Emily) Ropes
74593	The Baseball Boys of Lakeport: Or, The Winning Run	Edward Stratemeyer
74763	Lost Gip	Hesba Stretton