The *Novelties* Literary Dataset for Named Entity Recognition and Alias Resolution

Anonymous ACL submission

Abstract

We release *Novelties*, a literary English corpus for named entity recognition (NER) and alias resolution (AR). Compared to previous literary corpora, *Novelties* significantly improves over multiple aspects. First, our corpus contains 14 novels annotated in their *entirety*, which allows to study NER and AR at a more realistic scale. Second, *Novelties* includes *copyrighted* novels, which importantly allows to train and evaluate the performance of NER and AR systems on recent texts. While other corpora are limited to small portions of copyrighted texts, or outright exclude them, we take advantage of the recent `novelshare` alignment-based sharing technique to lawfully do so. Finally, *Novelties* is significantly *larger* than previous corpora in English, in terms of tokens. Through a series of experiments, we demonstrate the importance of these aspects for advancing the literary NLP landscape. We release *Novelties*, associated tools as well as our detailed annotation guidelines under a free license.

1 Introduction

Annotated corpora are crucial in the modern natural language processing (NLP) landscape. When it comes to the literary domain, while multiple datasets such as Litbank (Bamman et al., 2019, 2020) exist, they are often lacking in some key aspects. Specifically, most existing datasets only incorporate short excerpts and/or restrict themselves to public domain texts due to copyright limitations.

The reliance of datasets on short texts means the reported performance on annotated tasks can be deceptive. For example, previous research has shown that named entity recognition (NER) can benefit from additional context (Luoma and Pyysalo, 2020; Amalvy et al., 2023b,a) to improve performance. Meanwhile, coreference resolution is arguably harder at the level of an entire book (Guo et al., 2023; Gupta et al., 2024; Martinelli et al., 2025; Petersen-Frey et al., 2025). Computationally, some models are difficult to apply at such scale. Additionally, the task is more challenging and the performance of models significantly decrease at the level of a full novel. A push towards longer document is therefore necessary to fully study these tasks.

Meanwhile, the restriction to public domain texts results in an abundance of 19th century novels, to the detriment of recent novels, typically protected by copyright laws. That scarcity means that the performance of NLP systems on these texts is largely unknown, and that it is difficult to develop systems for them due to the lack of data. As highlighted by Amalvy et al. (2026), researchers find various ways to work around copyright limitations, such as including only short excerpts of copyrighted texts in the hope that this practice falls under fair use. Overall, these workarounds are not sufficient, and copyrighted texts remain rare in the landscape of literary datasets.

These issue motivates us to release *Novelties*, a new literary dataset for named entity recognition (NER) and alias recognition (AR). *Novelties* contains excerpts of novels as well as 14 full-length novels, addressing the problem of document scale. We also take advantage of the recent `novelshare` sharing technique (Amalvy et al., 2026) to lawfully include copyrighted texts in our dataset and share annotations publicly. Finally, *Novelties* is also significantly larger than existing datasets, being an order of magnitude larger than Litbank in number of tokens.

In order to demonstrate the importance of these specific characteristics of *Novelties*, we carry a series of experiments to highlight their influence. We discuss the significance of corpus scale in Section 6.2 and show its benefit to NER systems performance. In Section 6.3, we illustrate the value of long documents by highlighting the difficulty of the AR task at book scale. Finally, we establish the importance of incorporating recent copyrighted texts

in literary datasets by showing that NER performance on these texts is improved by the inclusion of recent material in training data.

Many of our choices when constituting *Novelties* result from our primary goal of creating a corpus suitable for the development systems that extract character networks (Labatut and Bost, 2019), such as Taggus (Canário et al., 2026). However, we think *Novelties* can be used for a broad range of applications in digital humanities. We release *Novelties* and our annotation guidelines freely. Additionally, we release related tools and source code to reproduce our experiments under the GPLv3 license[1].

2 Existing Datasets

Among NER and AR, the former is the most popular of the two, and a good number of literary NER corpora are freely available. In English, the *OWTO* corpus of Dekker et al. (2019) contains the first few chapters (often only the first) of 40 novels annotated in NER. To our knowledge, it is the only literary NER corpus to outright include copyrighted materials, as 20 of the included novels fall within this scope. One of the most notable literary corpus is *Litbank*, which contains excerpts of 100 annotated public domain novels (Bamman et al., 2019). A specificity of Litbank is that it is concerned with *nested NER*, a more complex formulation of the task where entities can contain other sub-entities. Litbank also has a French equivalent, called *fr-Litbank* (Mélanie-Becquet et al., 2024), but it is not as complete: it consists of excerpts of 18 novels plus two full novels. Still in French, Alrahabi et al. (2024) annotate excerpts of three novels. Later, the more recent corpus of Alrahabi et al. (2025) contains 7 19th century French novels (NCFN) annotated in their entirety. In Portuguese, PPORTAL_ner (Silva and Moro, 2024) is composed of 5,000 tokens from 25 different novels. Frontini et al. (2020) annotate excerpts of 400 words in the nine languages of the ELTeC corpus (Schöch et al., 2021). Finally, the *TravelWritings* corpus (Sprugnoli, 2018) is composed of 38 travelogues (i.e. accounts of an someone's traveling) from 1850 to 1940, and focuses on geographical entities.

Datasets for AR, meanwhile, are not as common. The French dataset of 7 novels of Alrahabi et al. (2025) mentioned above focuses on character AR, but neglects other types of entities. Other datasets are concerned with the more general task of coreference resolution. Litbank (Bamman et al., 2020) (and its French counterpart fr-Litbank (Mélanie-Becquet et al., 2024)) includes a coreference resolution layer, and remains a central reference when it comes to evaluation on literary texts. The FantasyCoref (Han et al., 2021) corpus covers excerpts from 213 fairy tales and Lewis Carroll's *Alice in Wonderland*. For German, The DROC corpus (Krug et al., 2018) contains semi-automatically coreference-annotated excerpts from 90 novels, while *FairyNet* (Schmidt et al., 2021) is composed of 40 annotated German fairy tales. A limitation of these datasets is that they focus on short texts (e.g. 2,000 tokens per novel for Litbank): other authors attempt to go further and propose corpora with full novels, albeit often with serious scale restrictions. Guo et al. (2023) annotate the entirety of George Orwell's *The Animal Farm*, but limit themselves to the top 20 characters. In German, Krug (2020) creates a small corpus with Theodore Fontane's novel *Effi Bries* and Ludwig Tieck's short fairy tale *Der blonde Eckbert*. Meanwhile, in Korean, the KoCoNovel dataset (Kim et al., 2024) is not limited by length as it is made up of 50 full-length novels.

As we can see from this section, existing literary datasets for NER and AR are lackluster in some aspects: 1) many of them do not include full texts (or only a few of them) and 2) to the best of our knowledge, only one of them (OWTO) includes copyrighted materials. Worse, none of them include full-length copyrighted texts. These important shortcomings therefore motivates us to release *Novelties*.

3 Scope and Constitution

Our focus is on two fundamental information extraction tasks: NER and AR. These are fundamental to character network extraction, and reflect our primary objective to develop resources to support the development of extraction tools such as Charnetto (Métrailler, 2023), Renard (Amalvy et al., 2024) or Taggus (Canário et al., 2026). However, it is worth stressing that their relevance extend largely beyond that context.

3.1 Named Entity Recognition

NER is a well-studied task, where the goal is to extract entity mentions along with their type, often by picking among a pre-determined set of classes.

[1] https://anonymous.4open.science/r/novelties-experiments-07F2

Corpus	Annotated	Material	Language	Tasks	©
DROC (Krug et al., 2018)	1 chapter	90 novels	de	CR	No
TravelWritings (Sprugnoli, 2018)	Full text	38 travelogues	en	NER	No
OWTO (Dekker et al., 2019)	~1st chapter	40 novels	en	NER	Yes
Litbank (Bamman et al., 2019)	2,000 tokens	100 novels	en	CR, NER	No
ELTeC (Frontini et al., 2020)	400 words	25 novels	en, fr +7	NER	No
HistoricNovels (Krug, 2020)	Full text	1 fairy tale + 1 novel	de	AR	No
FairyNet (Schmidt et al., 2021)	Full text	40 fairy tales	de	CR, NER	No
FantasyCoref (Han et al., 2021)	1700 tokens	213 fairy tales + 1 novel	en	CR	No
The Animal Farm (Guo et al., 2023)	Full text	1 novel	en	AR	No
OBVIL (Alrahabi et al., 2024)	Excerpts of chapters	3 novels	fr	NER	No
KoCoNovel (Kim et al., 2024)	Full text	50 novels	ko	CR	No
fr-Litbank (Mélanie-Becquet et al., 2024)	2 full texts + 18 $\times$ 10k tokens	20 novels	fr	CR, NER	No
PPORTAL_ner (Silva and Moro, 2024)	5,000 tokens	25 novels	pt	NER	No
NCFN (Alrahabi et al., 2025)	Full text	7 novels	fr	AR, NER	No
BookCoref (Martinelli et al., 2025)	Full text	3 gold + 50 silver novels	en	CR	No

Table 1: Summary of publicly available literary corpora annotated for NER, AR and coreference resolution (CR). Column *Annotated* indicates the length of the annotated excerpts, whereas *Material* shows where they come from. Column © indicates the presence of copyrighted material in the corpus.

A typical scheme involves persons (PER), locations (LOC), organizations (ORG), and miscellaneous (MISC), following the example of the CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) corpus.

We originally developed the NER layer of *Novelties* as a set of corrections and extensions to the OWTO corpus of Dekker et al. (2019). After obtaining a first stable version, we decided to write comprehensive annotations guidelines (Anonymous, 2024a), which led us to redefine the set of entities we annotate based on our main interest in character network extraction. We mainly modify the person (PER) class to include all characters (and therefore rename it CHR for *CHaRacter*), add a *group* (GRP) class to distinguish between single characters and groups of them and a more traditional *miscellaneous* (MSC) class. In details, we annotate the following classes:

- **Characters (CHR)**: individual entity with some form of sentience and agency in the plot. This includes humans but also some animals, mythical creatures, magical sentient weapons, robots...
- **Locations (LOC)**: mostly annotated as in many other corpora. We do not distinguish between natural (river, mountain, island...) and artificial (city, street, building, etc.) locations.
- **Organizations (ORG)**: explicitly organized institutions (e.g. government, company, etc.), as in many other corpora. We exclude informal groups such as families, *ethnonyms* (names given to the members of ethnic groups), and *demonyms* (names given to the inhabitants of some places).
- **Groups (GRP)** : informal gatherings of persons referred to under the same umbrella name, including family names, ethnonyms, and demonyms.
- **Miscellaneous (MSC)**: historical, cultural, commercial and sport events; unique objects and man-made products; meteorological events (eg. tempests); titles of intellectual or cultural works (eg. books, movies).

3.2 Alias Resolution

AR is the task of linking each alias in an extracted list of aliases to a single entity. In our case, we define an entity *alias* as the surface form of a mention already detected by NER. Therefore, AR can be thought as a limited version of the challenging coreference resolution task, that does not include certain types of mentions such as pronouns. We also make the important assumption that an alias serves to uniquely a single entity within the novel.

In addition to the linking task, we include in the AR task the question of creating a *canonical form* for each entity: a unique name that serves to precisely and uniquely identify it. The AR layer of *Novelties* is composed of the 14 novels that are entirely annotated for NER. We include all types of entities previously described (CHR, LOC, ORG, GRP and MSC).

As for NER, we write and release comprehensive annotation guidelines (Anonymous, 2024b) for AR.

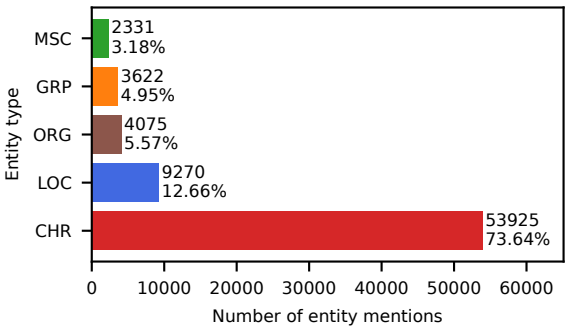

Figure 1: Distribution of entity mention types in the NER layer of *Novelties*.

Class	Kappa (all)	Kappa (annotated)	F1 score
CHR	96.99	94.23	96.63
LOC	88.78	87.11	86.57
ORG	84.43	82.86	84.06
GRP	87.38	86.73	88.19
MSC	81.13	80.35	82.86
All	94.20	85.17	92.58

Table 2: NER inter-annotator agreement metrics on different entity classes. We differentiate between the Kappa on all tokens or on annotated tokens only.

4 Description and Statistics

This section describes the NER and AR layers of *Novelties*. Table 7 in Appendix A provides some statistics for our entire corpus.

4.1 Named Entity Resolution

The NER layer of our corpus is composed of chapters from 48 different novels, 14 of which are entirely annotated. In total, this amounts to 599 chapters, for a total of 2,468,644 tokens and 73,186 entity mentions. *Novelties* is one order of magnitude larger than Litbank (Bamman et al., 2019) (210,532 tokens), and almost double the size of the French literary NER corpus from Alrahabi et al. (2025) (1,320,734 tokens). 24 of our 48 source novels are copyrighted, while the other 24 are from the public domain. We display the distribution of entity mention types in Figure 1. As can be seen, characters form the large majority (73.64%) of mentions, while other entity types are less common.

The *Novelties* dataset is entirely annotated by two annotators. To ensure high quality, we compute two inter-annotator agreement metrics: Cohen's Kappa (Cohen, 1960) and F1 score between annotators. We compute our metrics on a set of 5 chapters annotated by both annotators[2]. Since tokens that are not part of an entity mention (annotated as O in the BIO NER scheme) form a large majority, computing the Kappa on them would overestimate agreement (Brandsen et al., 2020). Therefore, we compute Cohen's Kappa on tokens that are part of a mention for at least one annotator to avoid that issue. Overall, we obtain a Kappa of 85.17 and a F1 score of 92.58, indicating excellent agreement. Table 2 shows more details and presents the agreement per class. We observe that it varies by class, with CHR displaying the highest agreement (96.63 F1), while the more imprecise MSC displays the lowest (82.86 F1).

4.2 Alias Resolution

Our alias resolution layer is constituted by all the novels that are fully annotated for NER, which comprises 14 full novels, 3,801 distinct entities and 5,888 aliases. To compute inter-annotator agreement, both annotators annotate Orwell's *1984* novel[3]. We focus on two different aspects of our alias resolution task: the unification of aliases and the definition of a canonical form. For the first aspect, we compute the Adjusted Rand index (ARI) (Rand, 1971). The original Rand index measures the proportion of pairs of names attached to the same or different entities by both annotators simultaneously. The adjusted index is a chance-corrected version of this measure. The second aspect is more difficult to evaluate, because there is not a single way to generate a correct canonical form. For reference, we report *Canonical Form Agreement*, the proportion of canonical forms that are almost identical, where we define "almost identical" as a Levenshtein ratio (Levenshtein, 1966) greater than 0.9.

Table 3 shows the agreement metrics of our two annotators on *1984*. ARI is high, especially for the CHR, GRP and MSC entity classes. It is, however, a bit lower for the LOC (71.54) and ORG (69.35) classes. For LOC entities, this is largely explained by the different handling of parts of locations by both annotators. For example, the first annotator differentiates geographical regions of Africa (*"Central Africa"*, *"North Africa"*, *"West African coast"*...),

[2]In the final version of the dataset, we include a version of these chapters where we resolved disagreements between annotators.

[3]As for NER, we include AR annotations with resolved annotator disagreements in the final version of the dataset.

Class	ARI	Canonical Form Agreement
CHR	90.05	81.25
ORG	69.35	70.69
LOC	71.54	74.23
GRP	100.0	37.04
MSC	97.63	59.18
All	83.86	68.46

Table 3: AR inter-annotator agreement metrics on different entity classes.

while the first annotator regroups these regions into a general *"Africa"* entity. For ORG entities, the lower agreement reflects the subtlety of judging the exact entity behind a specific form. The second annotator more often considers that two forms are a reference to the same entity without definite proof because that is likely (*"Eurasian General Staff"* and *"Eurasian High Command"*, *"Youth League"* and *"Youth Movement"*), while the first annotator disagrees. Other than these cases, we also observe a few errors by both annotators.

Canonical form agreement is relatively high, but very dependent on entity class. GRP agreement is especially low (37.04), not because of a strong disagreement but rather annotator's preference: for demonyms, the first annotator prefers adjectival forms (*"European people"*) while the second uses nouns (*"Europeans"*).

5 Sharing of Copyrighted Texts

One specific characteristic of *Novelties* is that the corpus contains *full copyrighted texts*. This addresses a longstanding issue where typical literary corpora such as Litbank (Bamman et al., 2019) or FantasyCoref (Han et al., 2021) are only composed of *public domain texts* because of legal restrictions. When developing models on these corpora, this dependency on freely available and often older texts can cause generalization issues (Lazaridou et al., 2021; Beelen et al., 2022). Furthermore, it generally restricts the data researchers have access to, meaning studies rarely target copyrighted data.

To get around copyright constraints, various strategies are currently employed as pointed by Amalvy et al. (2026). Generally, copyrighted materials are usually not shared, or shared only in parts. Since these choices have clear drawbacks, we adopt the recent novelshare sharing scheme (Amalvy et al., 2026). We provide only versions of our copyrighted novels where each token is represented by a truncated hash. To access our annotations, the corpus user must possess the materials themselves, and align their tokens with our annotations using the annotation algorithm provided by novelshare. Since the success of the alignment algorithm depends on the proximity of the user data to the original corpus data, we indicate editions used in our dataset in Appendix B.

6 Experiments

6.1 General Performance

In this section, we give a general idea of the performance of common inference systems when applied to *Novelties*.

6.1.1 Named Entity Recognition

We test two types of commonly used systems for NER: finetuned encoder-based models based on ModernBERT (Warner et al., 2025) and generative LLMs in a few-shots setting. In the case of ModernBERT-based models, we finetune them with a learning rate of $1e{-}5$ for a maximum of 2 epochs with early stopping as we empirically notice they are prone to overfitting. We experiment with ModernBERT-base (149M parameters) and ModernBERT-large (395M parameters). To obtain a score on every novel, we perform our evaluation 5-folds. For each fold, we use 10% of the training set as validation. In the case of generative models, we experiment with Llama-3.1-8B (Grattafiori et al., 2024). We did not test larger models due to the scale of our corpus rendering inference costs prohibitive. We implement a typed DSPy (Khattab et al., 2024) program, give the model one example per NER class and prompt it to output the entity mentions in the input passage of text. For all systems, we feed the input text in blocks of 3 sentences.

Table 4 shows the performance of the mentioned systems on *Novelties*. ModernBERT-based models perform the best (76.81 and 82.51 Micro-F1), with the large model beating the base one by almost 5 points. As rough points of comparison in the case of newswire datasets, Antoun et al. (2025) report that a model with a ModernBERT architecture achieves scores of approximately 92 F1 on the French TreeBank newswire dataset (Sagot et al., 2012), while Rueda et al. (2024) find that similar models such as LUKE (Yamada et al., 2020) obtain F1 scores around 94 F1 on the CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003).

Class	MBERT -base	MBERT -large	Llama -3.1-8B
CHR	82.00	87.88	48.25
LOC	74.67	78.61	50.69
ORG	44.17	53.71	21.86
GRP	61.24	65.26	8.75
MSC	37.01	45.47	5.56
All	76.81	82.51	38.67

Table 4: NER Micro-F1 score on *Novelties* for two variants of finetuned **ModernBERT** models and **Llama-3.1-8B** in a 5-shot setting.

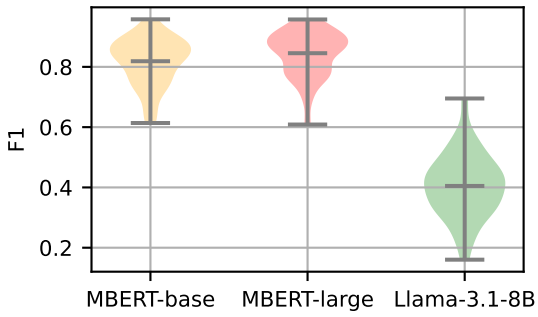

Figure 2: Distribution of NER Micro-F1 across *Novelties* novels for finetuned ModernBERT models and Llama-3.1-8B in a 5-shot setting. The middle bar of violins indicate the mean F1 over all novels.

The few-shots LLM setup clearly lags behind in terms of performance (38.67 F1), indicating the clear need of training a model for the task. We also observe an important difference of performance between entity classes: the CHR and LOC classes are the easiest, while ORG and MSC are more challenging. This difference generally reflects the difference in the distribution of entity mentions (Figure 1) and inter-annotator agreement metrics (Table 2), indicating a genuine difference of difficulty between classes.

We also display the performance of our surveyed systems across novels in Figure 2. This representation shows that performance varies widely over novels, with performance ranging from 60.87 to 95.73 F1 for ModernBERT-large. This confirms the earlier observation from Dekker et al. (2019) that the difficulty of literary NER is strongly dependent on the novel.

6.1.2 Alias Resolution

We benchmark the recent Maverick coreference system (Martinelli et al., 2024) on the AR layer of *Novelties*. While the Maverick model is able to

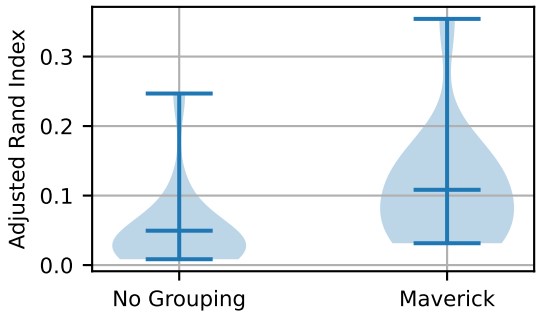

Figure 3: AR performance of our Maverick-based pipeline on *Novelties* novels compared to the "no grouping" clustering baseline. The middle bar of violins indicate the mean adjusted Rand index over all novels.

detect mentions, we only use it for linking entity mentions from the NER layer of *Novelties* since we do not want to perform full coreference resolution. We use the Maverick-large model trained on Litbank (Bamman et al., 2020). This pipeline only gives an upper bound of the performance of the system, as Litbank has excerpts of some of our novels and we feed it gold mentions at inference time. In order to meet our memory requirements, we divide our novels in chunks of 2,048 tokens and perform inference on these chunks. We then create document-level alias clusters by 1) linking together aliases that are in the same cluster at least once in the novel in a graph; and 2) extracting the connected components of this graph to form alias clusters. For clarity, we compare this clustering with a naive "no grouping" baseline where each alias is considered as a singleton cluster. We report the performance of our Maverick-based pipeline in terms of adjusted Rand index (Rand, 1971) on our 14 annotated novels in Figure 3.

As can be seen, performance of our Maverick pipeline is relatively low, with a mean adjusted Rand index (Rand, 1971) of 0.11[4]. As for coreference resolution (Guo et al., 2023; Gupta et al., 2024; Martinelli et al., 2025; Petersen-Frey et al., 2025), this confirms the difficulty of alias resolution at scale. As is the case for NER, performance also fluctuates a lot depending on the novel, suggesting novel-specific difficulties.

[4]Compared to typical coreference resolution datasets, our AR setting has important particularities: it comports a large majority of singletons (70.64%), and many small clusters. This leads to counter-intuitive results using some coreference metrics, and prevents a straightforward comparison with previous coreference resolution results. For reference, we include coreference metrics in Appendix C.

Class	MBERT -base full	MBERT -base 54%	MBERT -large full	MBERT -large 54%
CHR	82.00	78.75	87.88	86.49
LOC	74.67	68.56	78.61	76.50
ORG	44.17	37.75	53.71	48.07
GRP	61.24	56.39	65.26	63.36
MSC	37.01	23.85	45.47	40.85
All	76.81	72.60	82.51	80.67

Table 5: NER Micro-F1 score on *Novelties* for **ModernBERT-base** and **ModernBERT-large**, trained on either 1) the **full** version of the dataset or 2) **54%** of the dataset.

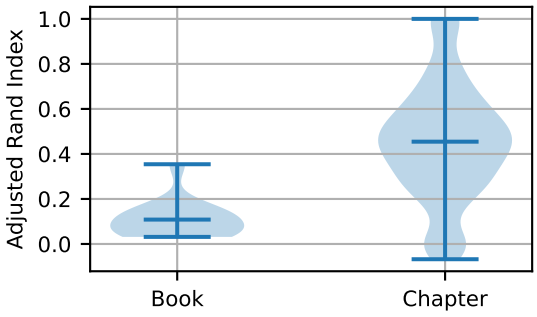

Figure 4: Performance of our Maverick-based AR pipeline on *Novelties* at the scale of entire books and chapters. Violins represent the performance distribution over books/chapters, their middle bar indicating the mean adjusted Rand index over the distribution.

6.2 Corpus Scale

The NER layer of *Novelties* is almost two times larger than the previous largest literary dataset by Alrahabi et al. (2025), and an order of magnitude larger than Litbank (Bamman et al., 2019). To understand the impact of this scale, we repeat our 5-folds NER experiment from Section 6.1.1 with encoder-based models, but compare the performance of models trained on the full size of the training set with models trained on a portion of it corresponding to the size of the dataset of Alrahabi et al. (2025), which amounts to 54% of the data. We present results in Table 5.

For both ModernBERT-base and ModernBERT-large, we observe a clear decrease of performance, noticeable for all entity classes. Additionally, this drop is far more important for the base model (-4.21 F1) compared to the large model (-1.84 F1). This result highlights the interest of having larger scale corpora to develop NER models.

6.3 Document Length

To discuss the interest of long document length, we compare the performance of our Maverick AR pipeline at the level of chapters and entire novels. As can be seen in Figure 4, AR is significantly easier at the chapter level (0.45 adjusted Rand index) than at the book level (0.11 adjusted Rand index), which confirms previous coreference resolution results (Guo et al., 2023; Gupta et al., 2024; Martinelli et al., 2025). This result demonstrates the importance of datasets with large document length for developing AR and coreference systems, and therefore the value of *Novelties* in the current literary NLP landscape.

6.4 Copyrighted Texts

To establish whether the fact that most previous literary datasets mostly avoid including copyrighted texts is problematic, we perform an additional experiment where we train encoder-based NER models as in Section 6.1.1 and report their performance on the copyrighted novels of our corpus. We compare between training on different data mixes: a *public domain* data mix, similar to most existing datasets, and a *public domain + copyrighted* one. To ensure a fair comparison between data mixes, we construct our training datasets so that they have approximately the same number of examples. In the case of the *public domain* data mix, we simply use all of the public domain novels of *Novelties* for training. For the *public domain + copyrighted* data mixes, we perform a 5-folds evaluation to obtain the performance on the entirety of our copyrighted data. For each fold, we use 80% of our copyrighted data as a training and validation set and complement with public domain data to match the number of training examples of the *public domain* data mix.

As can be seen in Table 6, we observe large performance improvement from using copyrighted data: +3.17 F1 for ModernBERT-base and +3.85 F1 for ModernBERT-large. These results show that including copyrighted texts in literary datasets is important, as not doing so is detrimental to performance.

7 Conclusion

In this article, we presented *Novelties*, a NER and AR English literary dataset. *Novelties* improves the state of literary NLP by being considerably larger

Class	MBERT-base public domain	MBERT-base public domain + ©	MBERT-large public domain	MBERT-large public domain + ©
CHR	79.61	83.10	82.87	86.45
LOC	64.49	66.59	66.05	70.58
ORG	46.70	50.86	51.72	54.38
GRP	40.48	43.90	43.94	51.76
MSC	27.84	30.62	26.94	40.99
All	72.70	75.87	75.77	79.62

Table 6: NER Micro-F1 score on the copyrighted novels of *Novelties* for **ModernBERT-base** and **ModernBERT-large**. Models are trained on either 1) all of the **public domain** data of *Novelties* or 2) a mix of copyrighted and public domain data of *Novelties* (**public domain + ©**).

than previous datasets, including long documents with 14 *entirely annotated* novels, and incorporating *copyrighted texts*, some of them in their entirety. Through empirical experimentation, we showed that these particularities have practical significance. The scale of our dataset means more data is available to train models, which leads to higher NER performance. We also validated the fact that AR at the scale of novels is significantly harder, making *Novelties* useful for the development and evaluation of AR systems. Finally, we established that the performance of NER models on recent copyrighted texts is improved by adding these types of texts in the training data, encouraging datasets towards lawfully including copyrighted materials.

Our primary goal in releasing this corpus is to support the development of character network extraction tools. However, we went beyond characters and annotated more types of entities. In the context of entity networks, our annotations can be used to develop systems that can extract new heterogeneous types of networks which may be used to study literary works with a different angle. As an example, Soni et al. (2023) extract spatial relationships between characters and places to explore literary questions such as character mobility, a use case that could be supported by character/location networks. We also think that NER and AR are tasks that are fundamental enough to be used in many other applications. For AR in particular, the difficulty of the problem at document scale warrants more research into that direction.

Limitations

Our experimental findings regarding the importance of dataset scale, document length and the inclusion of recent copyrighted texts are based on a few recent systems (namely, ModernBERT (Warner et al., 2025) for NER and Maverick (Martinelli et al., 2024) for AR). While they are representative of a subset of models used by the NLP community, findings may not completely generalize to other systems.

References

M. Alrahabi, A. Amalvy, V. Labatut, and P. Maurel. 2025. De l'annotation intégrale à l'analyse des réseaux de personnages : un modèle pour la ren dans les textes littéraires en français. *Revue TAL : traitement automatique des langues*, 66(1).

M. Alrahabi, C. Brando, F. Frontini, A. Provenier, R. Jalabert, M. Bordry, C. Koskas, and J. Gawley. 2024. Guide d'annotation manuelle d'entités nommées dans des corpus littéraires. Technical report, Campagne d'annotation OBVIL 2019–2021.

A. Amalvy, V. Labatut, X. Bost, and H.-H. Huang. 2026. Overcoming copyright barriers in corpus distribution through non-reversible hashing. In *64th Annual Meeting of the Association for Computational Linguistics*.

A. Amalvy, V. Labatut, and R. Dufour. 2023a. Learning to rank context for named entity recognition using a synthetic dataset. In *Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382.

A. Amalvy, V. Labatut, and R. Dufour. 2023b. The role of global and local context in named entity recognition. In *61st Annual Meeting of the Association for Computational Linguistics*, pages 714–722.

A. Amalvy, V. Labatut, and R. Dufour. 2024. Renard: A modular pipeline for extracting character networks from narrative texts. *Journal of Open Source Software*, 9(98):6574.

Anonymous. 2024a. Annotation guidelines for corpus Novelties: Part 1 — named entity recognition. Technical report.

Anonymous. 2024b. Annotation guidelines for corpus Novelties: Part 2 — alias resolution. Technical report.

W. Antoun, B. Sagot, and D. Seddah. 2025. ModernBERT or DeBERTaV3? examining architecture and data influence on transformer encoder models performance. In *14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, page 3061–3074.

A. Bagga and B. Baldwin. 1998. Algorithms for scoring coreference chains. In *1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

D. Bamman, O. Lewke, and A. Mansoor. 2020. An annotated dataset of coreference in English literature. In *12th Language Resources and Evaluation Conference*, pages 44–54.

D. Bamman, S. Popat, and S. Shen. 2019. An annotated dataset of literary entities. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2138–2144.

K. Beelen, J. Lawrence, D. C. S. Wilson, and D. Beavan. 2022. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship in the Humanities*, 38(1):1–22.

A. Brandsen, S. Verberne, M. Wansleeben, and K. Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *12th Language Resources and Evaluation Conference*, pages 4573–4577.

T. G. Canário, C. Duarte, F. L. Pinheiro, and J. L. M. Pereira. 2026. Taggus: an automated pipeline for the extraction of characters' social networks from portuguese fiction literature. *Social Network Analysis and Mining*, 16(1).

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

N. Dekker, T. Kuhn, and M. van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

F. Frontini, C. Brando, J. Byszuk, I. Galleron, D. Santos, and R. Stanković. 2020. Named entity recognition for distant reading in eltec. In *CLARIN Annual Conference*, pages 37–41.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, and 540 others. 2024. The llama 3 herd of models. *arXiv*, cs.AI:2407.21783.

Q. Guo, X. Hu, Y. Zhang, X. Qiu, and Z. Zhang. 2023. Dual cache for long document neural coreference resolution. In *61st Annual Meeting of the Association for Computational Linguistics*, pages 15272–15285.

T. Gupta, H. O. Hatzel, and C. Biemann. 2024. Coreference in long documents using hierarchical entity merging. In *8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–17.

S. Han, S. Seo, M. Kang, J. Kim, N. Choi, M. Song, and J. D. Choi. 2021. Fantasycoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *4th Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35.

O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *The Twelfth International Conference on Learning Representations*.

K. Kim, S. Lee, and S. Lee. 2024. KoCoNovel: Annotated dataset of character coreference in korean novels. *arXiv*, cs.CL:2404.01140.

M. Krug. 2020. *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Phd thesis, Julius Maximilians University Würzburg.

M. Krug, L. Weimer, I. Reger, L. Macharowsky, S. Feldhaus, F. Puppe, and F. Jannidis. 2018. Description of a corpus of character references in german novels - droc [deutsches roman corpus]. Technical Report 27, Georg-August-Universität Göttingen.

V. Labatut and X. Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys*, 52(5):89.

A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liška, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kocisky, S. Ruder, D. Yogatama, K. Cao, S. Young, and P. Blunsom. 2021. Mind the gap: assessing temporal generalization in neural language models. In *35th International Conference on Neural Information Processing Systems*, pages 29348–29363.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

X. Luo. 2005. On coreference resolution performance metrics. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

J. Luoma and S. Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *28th International Conference on Computational Linguistics*, pages 904–914.

G. Martinelli, E. Barba, and R. Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), page 13380–13394.

G. Martinelli, T. Bonomo, P. Huguet Cabot, and R. Navigli. 2025. Bookcoref: Coreference resolution at book scale. In *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 24526–24544.

N. S. Moosavi and M. Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

F. Mélanie-Becquet, J. Barré, O. Seminck, C. Plancq, M. Naguib, M. Pastor, and T. Poibeau. 2024. BookNLP-fr, the French versant of BookNLP. a tailored pipeline for 19th and 20th century French literature. *Journal of Computational Literary Studies*, 3(1):1–34.

C. Métrailler. 2023. Charnetto.

F Petersen-Frey, H. O. Hatzel, and C. Biemann. 2025. Efficient and effective coreference resolution for german. In *21st Conference on Natural Language Processing (KONVENS 2025)*, pages 128–137.

W. M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

M. Recasens and E. Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17:485–510.

A. Rueda, E. Alvarez-Mellado, and C. Lignos. 2024. CoNLL#: Fine-grained error analysis and a corrected test set for CoNLL-03 English. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3718–3728.

B. Sagot, M. Richard, and R. Stern. 2012. Annotation référentielle du corpus arboré de Paris 7 en entités nommées (referential named entity annotation of the Paris 7 French TreeBank) [in French]. In *Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 535–542.

D. Schmidt, A. Zehe, J. Lorenzen, L. Sergel, S. Düker, M. Krug, and F. Puppe. 2021. The fairynet corpus - character networks for german fairy tales. In *5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, page 49–56.

C. Schöch, R. Patras, T. Erjavec, and D. Santos. 2021. Creating the european literary text collection : Challenges and perspectives. *Modern Languages Open*, 1(25):1–19.

M. O. Silva and M. M. Moro. 2024. PPORTAL_ner: An annotated corpus of portuguese literary entities. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12927–12937.

S. Soni, A. Sihra, E. F. Evans, M. Wilkens, and D. Bamman. 2023. Grounding characters and places in narrative texts. In *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 11723–11736.

R. Sprugnoli. 2018. Arretium or Arezzo? a neural approach to the identification of place names. In *5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 361–366.

E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *7th Conference on Natural Language Learning*, pages 142–147.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *6th Message Understanding Conference*.

B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, G. T. Adams, J. Howard, and I. Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Conference on Empirical Methods in Natural Language Processing*, pages 6442–6454.

A Detailed Statistics

We present detailed statistics for our corpus in Table 7.

B Editions

We use the `novelshare` sharing scheme (Amalvy et al., 2026) in order to lawfully share the copyrighted portion of *Novelties*. Since the performance of the `novelshare` alignment algorithm depends on the proximity of the user data to the shared data, we indicate the editions of *Novelties* copyrighted novels in Table 8 when they are available to us.

C Coreference Metrics for Alias Resolution

In this section, we report the results of the AR systems we survey in Section 6.1.2 in terms of traditional coreference metrics (see Table 9): MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) $CEAF_{\phi_4}$ (Luo, 2005), BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016).

Novel	Chapters	Tokens	Entity Mentions	Entity Aliases	Entities	Full	US ©
1984	23/23	116,919	2,269	277	207	yes	yes
A Game of Thrones	1/73	3,763	181	-	-	no	yes
Alice in Wonderland	3/12	7,372	198	-	-	no	no
Assassin's Apprentice	1/25	7,865	149	-	-	no	yes
A Study in Scarlet	1/14	3,396	62	-	-	no	no
Bel Ami	18/18	130,987	2,653	571	395	yes	no
Beyond the Shadows	1/100	1,215	75	-	-	no	yes
Black Prism	1/95	4,203	124	-	-	no	yes
Brave New World	18/18	80,131	2,420	534	393	yes	yes
David Copperfield	1/64	5,645	106	-	-	no	no
Dracula	1/27	6,599	103	-	-	no	no
Elantris	1/65	5,224	203	-	-	no	yes
Emma	2/75	5,902	174	-	-	no	no
Eugénie Grandet	14/14	80,108	2,513	368	223	yes	no
Frankenstein	1/28	8,248	74	-	-	no	no
Gardens of the Moon	1/26	14,371	395	-	-	no	yes
Germinal	40/40	208,973	4,935	398	277	yes	no
Harry Potter	1/17	5,668	232	-	-	no	yes
Huckleberry Finn	2/43	4,566	90	-	-	no	no
Madame Bovary	35/35	138,618	3,258	692	501	yes	no
Magician: Master	1/34	8,868	318	-	-	no	yes
Mistborn	1/40	3,677	172	-	-	no	yes
Moby Dick	136/136	250,153	5,260	1,381	1,017	yes	no
Oliver Twist	3/53	9,791	214	-	-	no	no
Pride and Prejudice	4/61	5,307	156	-	-	no	no
Prince of Thorns	1/49	1,198	35	-	-	no	yes
Storm Front	1/27	2,707	36	-	-	no	yes
Dr Jekyll and Mr Hyde	1/10	2,832	29	-	-	no	no
The Black Company	7/7	113,303	4,989	241	170	yes	yes
The Blade Itself	46/46	240,763	7,799	494	251	yes	yes
The Call of the Wild	1/7	4,364	115	-	-	no	no
The Colour of Magic	6/6	78,765	2,570	371	214	yes	yes
The Count of Monte Cristo	1/117	4,071	146	-	-	no	no
The Fellowship of The Ring	1/23	12,032	473	-	-	no	yes
The Gunslinger	3/65	3,068	76	-	-	no	no
The Hunchback of Notre-Dame	59/59	219,481	6,512	1,928	1,209	yes	no
The Lies of Locke Lamora	1/33	14,574	576	-	-	no	yes
The Light Fantastic	7/7	78,911	2,682	333	194	yes	yes
The Name of the Wind	1/94	7,658	232	-	-	no	yes
The Painted Man	1/32	12,162	599	-	-	no	yes
The Red and The Black	75/75	216,699	6,847	1,076	632	yes	no
The Three Musketeers	68/68	289,672	11,313	855	564	yes	no
The Way of Kings	1/87	8,957	357	-	-	no	yes
The Way we Live Now	2/100	8,707	157	-	-	no	no
The Wheel of Time	1/53	8,592	357	-	-	no	yes
Tinker Tailor Soldier Spy	1/39	5,955	5,955	-	-	no	yes
Ulysses	1/18	8,774	407	-	-	no	no
Vanity Fair	1/67	7,830	302	-	-	no	no
Total	599	2,468,644	73,186	9,519	6,247	14	24

Table 7: Summary statistics of *Novelties*. Column *US* © indicates whether a novel is copyrighted in the USA. We only indicate number of entity and entity aliases for full novels, since we don't annotate incomplete novels for AR.

Novel	ISBN
1984	9781405879439
A Game of Thrones	9780307292094
Assassin's Apprentice	9780553897487
Beyond the Shadows	9781607513513
Brave New World	9780795300028
Elantris	9780765383105
Gardens of the Moon	9788498003178
Harry Potter	9781781103685
Magician Master	9780007466863
Mistborn	9788374805537
Prince of Thorns	9781101543290
Storm Front	9781101128657
The Black Company	N/A
The Black Prism	9782352945260
The Blade Itself	9780575079793
The Colour of Magic	9780061796845
The Fellowship of the Ring	9780547952017
The Lies of Locke Lamora	9780575079755
The Light Fantastic	N/A
The Name of the Wind	9782352949152
The Painted Man	9780007518616
The Way of Kings	9780765326355
The Wheel of Time	9781857230765
Tinker Taylor Soldier Spy	N/A

Table 8: Copyrighted novels and editions used in *Novelties*. The information is unavailable to us for three novels. *Tinker Tailor Soldier Spy* originates from the OWTO corpus (Dekker et al., 2019), and its original ISBN is not available. For the two remaining novels, the epub files used for annotation did not include the information.

We only report these results for reference only, as our AR setting comports a large majority of singletons and small clusters. This leads to counterintuitive results for some metrics, and renders comparison with coreference results from the literature difficult. As a very rough comparison, a Maverick-based model obtains 94.3 MUC, 55.3 B^3 and 33.4 $CEAF_{\phi_4}$ F1 at the novel scale on $BookCoref_{gold}$ (Martinelli et al., 2025).

D Few-shot NER DSPy Program

To perform LLM-based NER inference in Section 6.1.1, we use a typed DSPy program. We directly use a shortened version of entity types definition in our guidelines to the prompt. Our program is provided in Figure 5.

System	MUC	B^3	$CEAF_{\phi_4}$	BLANC	LEA
No Grouping	0.00	79.87	69.65	49.92	51.06
Maverick	49.49	80.15	69.17	54.54	59.70

Table 9: Coreference metrics F1 score of surveyed AR systems on the AR layer of *Novelties*.

```
class EntityExtraction(dspy.Signature):
    """
    Extract named entities from a text. The 5 possible entity classes are:
    - CHR (Character): characters in a literary sense. Includes names of people, but
        also some non-humans such as animals, robots, magical creatures...
    - LOC (Location): Name of places. Includes natural (river, mountain...) or
        artificial (city, street...) places.
    - ORG (Organization): Organized institutions such as government, companies...
        Excludes informal groups such as families, ethnonyms and demonyms.
    - GRP (Group): Informal groups of characters such as families, ethnonyms or
        demonyms.
    - MSC (Miscellaneous): Heterogeneous category gathering praxonyms (historical,
        cultural, commercial or sport events), ergonyms (objects and man-made
        products) provided they are unique, phenonyms (meteorological events such as
         tempests), and titles of intellectual or cultural works (books, movies...).
    """

    class DSPyNEREntity(TypedDict):
        ner_class: Literal["CHR", "ORG", "LOC", "GRP", "MSC"]
        tokens: list[str]

    tokens: list[str] = dspy.InputField(desc="tokenized␣text")
    entities: list[DSPyNEREntity] = dspy.OutputField(desc="list␣of␣extracted␣
        entities")
```

Figure 5: Typed DSPy program used to perform the LLM-based NER inference in Section 6.1.1.