jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

BookNLP-fr, the French Versant of BookNLP

A Tailored Pipeline for 19th and 20th Century French Literature

Frédérique Mélanie-Becquet[1]
Jean Barré[1]
Olga Seminck[1]
Clément Plancq[2]
Marco Naguib[3]
Martial Pastor[4]
Thierry Poibeau[1]

1. Lattice UMR 8094, École Normale Supérieure – PSL – CNRS – Université Sorbonne Nouvelle, Montrouge, France.
2. MSH Val de Loire UAR 3501, CNRS – Université de Tours – Université d'Orléans, Tours, France.
3. LISN, Université Paris-Saclay and CNRS, Orsay, France.
4. Centre for Language Studies, Radboud University, Nijmegen, The Netherlands.

Citation
Frédérique Mélanie-Becquet, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau (2024). "BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature". In: *Journal of Computational Literary Studies* 3 (1). 10.48694/jcls.3924

Date published 2024-11-21
Date accepted 2024-09-29
Date received 2024-01-25

Keywords
natural language processing, computational literary studies, French literature, coreference resolution, entity recognition, subgenre classification

Reviewers
Glenn Roe, Laurent Romary

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 3rd Annual Conference of Computational Literary Studies at Vienna, Austria, in June 2024.

Abstract. This paper presents BookNLP-fr: the adaptation to French of BookNLP, an existing NLP pipeline tailored for literary texts in English. We provide an overview of the challenges involved in the adaptation of such a pipeline to a new language: from the challenges related to data annotation up to the development of specialized modules of entity recognition and coreference. Moving beyond the technical aspects, we explore practical applications of BookNLP-fr with a canonical task for computational literary studies: subgenre classification. We show that BookNLP-fr provides more relevant and – even more importantly – more interpretable features to perform automatic subgenre classification than the traditional *bag-of-words* approach. BookNLP-fr makes NLP techniques available to a larger public and constitutes a new toolkit to process large numbers of digitized books in French. This allows the field to gain a deeper literary understanding through the practice of distant reading.

1. Introduction

The domain known as Computational Humanities has recently emerged with the availability of large corpora of literary texts in digitized format, and of transformer-based language models that are quick, robust, and (generally) accurate (e.g., Devlin et al. 2019; Touvron et al. 2023). This situation opened up new opportunities for exploration and analysis. For French, the collection *Literary Fictions of Gallica* (Langlais 2021) includes 19,240 public domain documents from the digital platform of the French National Library, enabling researchers to navigate the wide diversity of literature with unprecedented ease.

The sheer volume of digitized texts presents a unique set of challenges. Traditional methods of literary analysis and interpretation are insufficient when confronted with such vast corpora. It is no longer feasible for individuals to manually analyze the entirety of these collections through close reading. This shift in scale necessitates the development of innovative tools and technologies, particularly Natural Language Processing (NLP). These tools are essential for extracting meaningful insights from digital corpora. They can illuminate patterns, trends, and connections that would be impractical or impossible for humans to discern within the vast amount of text data. This new technical paradigm opens up the possibility of conducting research through distant reading (Moretti 2000; Underwood 2019), enabling scholars to zoom in and out from the literary past, facilitating a more profound comprehension of trends and patterns that delineate the evolution of literature. The knowledge embedded in these digitized literary corpora is crucial not only for literary scholars but also for those interested in cultural analytics, defined as "the analysis of massive cultural datasets and flows using computational and visualization techniques" by Manovich (2016), or more practical applications for example the automatic production of book summaries for catalogs (Zhang et al. 2019). The evolution of literature is intricately tied to broader shifts in society, and digitized texts offer a unique opportunity to study these transformations.

To make the analysis of such large corpora possible, BookNLP (Bamman 2021) has been proposed as a specialized software solution adapted to literary texts. It includes the analysis of entities, coreference, events, and quotations within textual data. Originally conceived at the University of California, Berkeley in 2014 by David Bamman and his team, BookNLP has undergone continuous enhancements in line with the latest advancements in NLP. Notably, it has embraced emerging technologies such as integrated embeddings of Large Language Models (LLMs), more specifically BERT (Devlin et al. 2019) in early 2020.

The ongoing evolution of BookNLP extends beyond its initial scope, as efforts are underway to expand its applicability to five additional languages through the Multilingual BookNLP project (Bamman 2020). However, it is worth noting that French is not included in this extension. In response to this gap, it was decided in 2021, in coordination with Berkeley, to develop a dedicated French version of BookNLP. The goal is that researchers working with French literary data have access to the basic tools required for the structured analysis of fiction. This paper thus presents the French BookNLP project, the related annotated corpus, and the pieces of software defined within the project, as well as a specific study that illustrates how BookNLP can be used for literary studies.

The structure of the paper is as follows: We start with a literature review in which we specify NLP tools and techniques that are of particular interest in a framework for distant reading (section 2). Special attention will be given to the results of the English BookNLP project (subsection 2.2). In section 3, we provide a detailed description of how we elaborated the pipeline of BookNLP-fr: the training data, the annotation process, and the software development. In section 4, we give the evaluation scores of our pipeline on the subtasks of entity recognition and coreference resolution. Then, we present a case study where we used BookNLP-fr for the classification of literary genre (section 5). We conclude this article with a discussion of how the use of computational methods and the framework of distant reading with imperfect annotations affects the field of literary studies (subsection 6.1) and its perspectives in the era of *LLMs* (subsection 6.2), and finally summarize the paper in section 7.

2. Literature Review

2.1 Computational Methods Applied to Literary Text Analysis

Statistical methods have been used extensively in literary text analysis to identify patterns and trends in large amounts of textual data. Different pieces of software are available for this, for example, quanteda (Benoit et al. 2018), stylo (Eder et al. 2016), TidyText (Silge and Robinson 2017), or Voyant Tools (Rockwell and Sinclair 2016), to cite the most famous. They are available 'off the shelf', which means that they can be used directly by scholars and researchers to analyze texts. These tools can handle raw text directly, or after basic NLP processes such as lemmatization, part-of-speech tagging, or other kinds of annotation. They offer various visualizations to interpret the texts, such as dendrograms to represent the 'distance' between various books in a corpus, or charts to show what type of vocabulary is typical to one author as opposed to another.

There are clear benefits in using statistical methods to analyze literary texts, such as the ability to process and analyze large amounts of data quickly and efficiently, to identify patterns and trends that might not be apparent through traditional close reading methods, and to generate new research questions and hypotheses. NLP is needed to better represent the content of the text, i.e., what the text says behind the words used. NLP techniques can be used to annotate literary texts by providing syntactic and semantic annotations. NLP has become an increasingly important tool in the field of literary studies, providing new methods for analyzing and interpreting literary texts. NLP tools (e.g., NLTK (Bird et al. 2019) or Stanford CoreNLP Toolkit (Manning et al. 2014)) have been used to perform a wide range of tasks, including part-of-speech tagging, syntactic analysis, named entity recognition, etc. In the following paragraphs, we will specify the linguistic analyses available by the BookNLP pipeline: entity recognition, coreference resolution, event recognition and quotation detection. The tools mentioned in the paragraph above do not provide these types of semantic analyses, but only use morphological and grammatical linguistic analyses. BookNLP thus occupies a special niche and provides more semantically oriented annotations.

Entity Recognition. Entity recognition, along with coreference resolution, is of prominent importance, since it makes it possible to track characters, their actions, and their relationships over time. Named entity recognition (NER) is a well-established task in NLP, referring to the recognition of persons, locations, companies, other institutions, etc. (Maynard et al. 2017). NER systems exist for a wide array of languages (Emelyanov and Artemova 2019), with generally good performance, depending of course on the nature of the document to be analyzed and of the gap between training data and target data. Recognizing mentions referring to characters in a novel shares many features with NER, but is more varied (not all characters have a name, and a character can correspond to an animal, for example). Locations are also of the utmost importance to track the movements of characters (Ryan et al. 2016), but also to detect events. Note that performance may vary greatly depending on the nature of the text and of the entities to be recognized, for example, in the novel *Les Mystères de Paris*, written between 1842 and 1843 by Eugène Sue, most of the characters have names that are similar to noun phrases, such as "la Goualeuse" (meaning *the Street Singer*) or "le Chourineur" (meaning *the Stabber*). Also science fiction, which is full of non-classical proper nouns, can be very challenging for the task (Dekker et al. 2019). A module able to predict, or at least estimate, performance from cues gathered in the text would be useful to process large collections of novels.

Coreference Resolution. In the sense of linking together all the mentions in the text of a given character, although the task can involve all kinds of names or even nouns, coreference resolution is challenging by nature. There is a long tradition of research in coreference resolution in NLP, and modules exist for different languages, with various levels of performance (Poesio et al. 2023). The quality of the different systems is still increasing (through end-to-end models (Lee et al. 2017) and then transformer-based language models (Joshi et al. 2019)), and coreference resolution remains a very active field of research in NLP. The task is more challenging for French or Russian than for English, since the 'it' pronoun limits ambiguity in English, whereas in French all nouns are masculine or feminine, not only human beings, and are referred to with third-person pronouns. For instance, in *"Marie veut qu'on lave la voiture, **elle** est sale."* (*"Marie wants that we wash the car, **it** is dirty."*), *elle* refers to *the car* but could theoretically also refer to *Marie*; from a human point of view there is no ambiguity in this sentence, but the analysis requires semantic information. When applied in literary studies, automatic coreference systems often break long coreference chains due to the fact that they use a fixed-sized sliding window. If a given character does not appear during a certain period of time (i.e., a certain number of pages), it makes it harder to retrieve its antecedent. Literature provides a good testbed for the coreference task, since novels are long, real, and complex texts on which performance can (and should) still improve a lot.

Event Recognition. Event recognition involves the automated identification and extraction of verbs and, more rarely, nouns referring to events. The task is difficult in that there is no clear definition of what an event is, and other features interact with the definition (among others: negation, adverbials, and modals), and not all occurrences of verbs should be annotated (e.g., in *"I like to play tennis"*, *play* is an infinitive that refers to something I like, but it is generally considered that there is no event *per se* in the sentence). As for literary texts, there have been initiatives to annotate events (Sims et al. 2019), but most verbs and even some nouns can refer to events (Hogenboom et al. 2016; Sprugnoli and Tonelli 2016), which may lead to a too fine-grained annotation. There is thus a need to redefine the task and provide an intermediate level of annotation, between isolated events and the novel as a whole (Lotman 1977; Schmid 2010a,b), but also higher level annotations (like the notion of scene) have proven difficult to formalize, leading to very low accuracy in practical experiments (Zehe et al. 2021).

Quotation recognition plays a crucial role in enhancing the understanding of textual content by identifying and isolating direct speech instances. This feature is instrumental in extracting and preserving the spoken words of characters, enabling a fine-grained analysis of dialogue patterns and character interactions (Durandard et al. 2023; van Cranenburgh and van den Berg 2023). A crucial but complex part of the task consists in establishing which character is at the origin of a given utterance. A recent study has shown that performance on this task is still rather low and would need to be improve to be really usable in operational contexts (Vishnubhotla et al. 2023).

2.2 The BookNLP Project

BookNLP is a set of NLP modules designed specifically for the analysis of novels and other literary prose texts. Developed by David Bamman and colleagues at the University of Berkeley (Bamman 2021; Bamman et al. 2014), BookNLP employs a combination of machine learning and linguistic analysis techniques to extract information from text and perform tasks such as character recognition, coreference resolution, event recognition, and quotation extraction.[1] The annotated files that are available for training constitute the LitBank corpus (Bamman et al. 2020, 2019). This corpus is publicly available[2], which makes it possible to regularly retrain the system as NLP continues to evolve rapidly (especially LLMs).

Entity Recognition. One of the primary tasks of BookNLP is entity recognition, more specifically the recognition of characters, locations, and vehicles, showing the focus on the actions of characters. This information is used, i.a., to study how mobile protagonists are and what kind of space male and female characters occupy (Soni et al. 2023). Character recognition is often coupled with other information (such as gender, attributes, relations between characters) that can be useful for sub-stream tasks.

Coreference Resolution. In the context of literature, coreference resolution often involves resolving pronouns and other referring expressions to specific characters or entities. BookNLP employs advanced linguistic analysis to identify and link references to the same entity, and the extra knowledge provided by LLMs is especially useful for this task.

Event Recognition. Another essential task performed by BookNLP is Event recognition. It should be crucial for analyzing the development of the storyline and identifying key plot points, but the huge number of verbs supporting actions makes the annotation too prolific and not adapted to specific needs. The proper annotation of negations, adverbs, and modals is also an open problem. This is why event recognition has not been addressed as a priority in the context of the Multilingual BookNLP Project, which rather focus on entity recognition and coreference resolution.

Quotation Extraction. BookNLP is equipped with the capability to extract quotations from a text. This involves identifying and isolating the direct speech or quoted passages within the literary work. Accurate quotation extraction is vital for understanding character dialogue, the intentions of characters, and developing further analyses. However, quotation recognition without speaker attribution is not so useful, and as we have seen, speaker attribution remains an open question, as accuracy for the task remains low (Vishnubhotla et al. 2023).

The application of BookNLP for the analysis of novels and other literary works aims at providing a deeper understanding of narrative structures, character dynamics, and thematic elements in novels (Piper et al. 2021). The different modules are intended

[1] Note that BookNLP suite currently is based upon BERT (e.g., Devlin et al. 2019), but this could evolve as better language models continue to appear.
[2] See: `https://paperswithcode.com/dataset/litbank`.

to assist researchers in literary analysis, but also in digital humanities and cultural analytics.

3. French BookNLP

The French BookNLP project endeavors to construct a robust NLP pipeline specifically tailored for the comprehensive analysis of extensive French literary corpora of the 19th and 20th century. The ongoing Multilingual BookNLP project (Bamman 2020), coordinated by Berkeley, seeks to update the initial pipeline (Bamman et al. 2014) and extend its capabilities to encompass four additional languages (Spanish, German, Russian and Japanese). In alignment with this initiative – even though we are not part of the Multilingual BookNLP project itself, in the sense that we are independent of the research grant received by the Berkeley team – we are actively engaged in the development of the necessary linguistic resources for the French language. Our collaboration with the Berkeley project ensures a coordinated approach to this expansion, e.g., by sharing similar annotations and visualization tools.

In line with the Multilingual BookNLP project, we focus mainly on entity recognition and coreference resolution. We have seen in the previous sections that annotating events entails a number of problems and may be too general, and thus not useful, if it is not done with a specific goal in mind (which may entail some domain-specific annotations, with adapted categories, for example). We have also seen that quotation recognition without a proper speaker attribution algorithm is not really useful for similar reasons, but that speaker attribution remains an open problem (Brunner et al. 2020). In the following, we will thus not address these two tasks (namely event and quotation recognition) for further investigation and concentrate on entity recognition and coreference resolution.

3.1 The Training Corpus and the Democrat Project

The Democrat Project (hereafter just *Democrat*), led by Frédéric Landragin (2016, 2021), funded by the French National Research Agency (ANR) and completed in 2020, aimed to develop an annotated corpus at the level of coreference chains in French. Before the Democrat, no such corpus existed. One of the fundamental aspects of Democrat was the annotation of long texts, in contrast to, e.g., the Ontonotes corpus (Weischedel et al. 2013), which serves as a standard for English but is predominantly composed of short texts. Additionally, the Democrat project aimed to annotate a wide variety of text types, including novel chapters, short stories, journalistic pieces, legal documents, encyclopedic entries, technical texts, and more. It also had a diachronic dimension, spanning from medieval French to contemporary French.

For the needs of the BookNLP-fr project, we focused on annotations related to novels and selected texts spanning from the early 19th century to the early 20th century. Before this period, French is more prone to variation, and for the more recent period, texts are not freely shareable due to copyright issues. Lastly, to keep the annotation task manageable, each text in the Democrat corpus is actually composed of a 10,000-word excerpt (leaving us with 184,137 tokens). In addition to this selection from Democrat, we added two short stories by Balzac, which account for 45,238 tokens. Information about these texts and those from Democrat can be found in Table 1.

Year	Author	Title	Source
1830	Honoré de Balzac	*La maison du chat qui pelote*	Full Text
1830	Honoré de Balzac	*Sarrasine*	Democrat 10 K
1836	Théophile Gautier	*La morte amoureuse*	Democrat 10 K
1837	Honoré de Balzac	*La maison Nucingen*	Full Text
1841	George Sand	*Pauline*	Democrat 10 K
1856	Victor Cousin	*Madame de Hautefort*	Democrat 10 K
1863	Théophile Gautier	*Le capitaine Fracasse*	Democrat 10 K
1873	Émile Zola	*Le ventre de Paris*	Democrat 10 K
1881	Gustave Flaubert	*Bouvard et Pécuchet*	Democrat 10 K
1882-1883	Guy de Maupassant	*Mademoiselle Fifi, nouveaux contes* (1)	Democrat 10 K
1882-1883	Guy de Maupassant	*Mademoiselle Fifi, nouveaux contes* (2)	Democrat 10 K
1882-1883	Guy de Maupassant	*Mademoiselle Fifi, nouveaux contes* (3)	Democrat 10 K
1901	Lucie Achard	*Rosalie de Constant, sa famille et ses amis*	Democrat 10 K
1903	Laure Conan	*Élisabeth Seton*	Democrat 10 K
1904-1912	Romain Rolland	*Jean-Christophe* (1)	Democrat 10 K
1904-1912	Romain Rolland	*Jean-Christophe* (2)	Democrat 10 K
1917	Adèle Bourgeois	*Némoville*	Democrat 10 K
1923	Raymond Radiguet	*Le diable au corps*	Democrat 10 K
1926	Marguerite Audoux	*De la ville au moulin*	Democrat 10 K
1937	Marguerite Audoux	*Douce Lumière*	Democrat 10 K

Table 1: The texts in the BookNLP-fr corpus.

3.2 Data Preparation and Annotation

In the scope of the Democrat project, annotations have been applied to all types of coreference. However, for the BookNLP-fr project, our specific focus lies within a subset of these coreferences, corresponding to certain types of entities: persons (PER), facilities (FAC), geo-political entities (GPE), locations (LOC), vehicles (VEH), organizations (ORG), and denotations of time (TIME). Definitions from all these categories except for time are adapted from Bamman et al. (2019).

PER. According to Bamman et al. (2019, 2139): "By person we describe a single person indicated by a proper name (**Tom Saywer**) or common entity (**the boy**); or set of people, such as **her daughters** and **the Ashburnhams**". Examples of PER from our corpus can be found in example (1) and example (2)[3]:

(1) a. une de ces gentilhommières si communes en Gascogne, et que **les villageois** décorent du nom de château Le Capitaine Fracasse
    b. one of those manors so common in Gascogne, and that **the villagers** decorated by the name of the castle of Captain Fracasse

(2) a. **Madame François**, adossée à une planchette contre **ses** légumes
    b. **Madame François**, who leaning on a board next to **her** vegetables

FAC. We follow Bamman et al. (2019, 2139)'s definition: "For our purposes, a facility is defined as a 'functional, primarily man-made structure' designed for human habitation (buildings, museums), storage (barns, parking garages), transportation infrastructure

[3] Note that PER mentions are split into three parts to enable more fine-grained analyses, including proper nouns (PROP), common phrases (NOM), and pronouns (PRON). Pronouns account for the majority of mentions, specifically 59%, 32%, and 9%, respectively.

(streets, highways), and maintained outdoor spaces (gardens). We treat rooms and closets within a house as the smallest possible facility". See example (3):

(3) a. **Le chemin** qui menait de **la route** à **l'habitation** s'était réduit, par l'envahissement de la mousse et des végétations parasites
b. **The path** that led to **the road** to **the dwelling** was narrowed by the invasion of moss and parasitic vegetation

GPE. We follow Bamman et al. (2019, 2139)'s guidelines for this category: "Geopolitical entities are single units that contain a population, government, physical location, and political boundaries.". See example (4):

(4) a. Échappé de **Cayenne**, où les journées de décembre l'avaient jeté, rôdant depuis deux ans dans **la Guyane hollandaise**, avec l'envie folle du retour et la peur de la police impériale, il avait enfin devant lui **la chère grande ville**, tant regrettée, tant désirée.
b. Escaped from **Cayenne**, where the December days had thrown him, erring since two years in **Dutch Guyane**, with a crazy desire to return and fear of the imperial police, he finally had before him the **dear big city**, so much regretted and desired.

LOC. As opposed to GPEs, Bamman et al. (2019, 2139) define locations as "entities with physicality but without political organization [...] such as **the sea**, **the river**, **the country**, **the valley**, **the woods**, and **the forest**". See the example (5) and the example (6) from our corpus:

(5) a. des moellons effrités aux pernicieuses influences de **la lune**
b. crumbling rubble masonry under the pernicious influences of **the moon**

(6) a. Poussez-moi ça dans **le ruisseau** !
b. Push this into **the stream** !

VEH: The definition for a vehicle is a *"physical device primarily designed to move an object from one location to another"* (Bamman et al. 2019). An example from our corpus:

(7) a. anciennement **des voitures** avaient passé par là
b. before, **carriages** had passed there

ORG. According to Bamman et al. (2019, 2139), "[o]rganizations are defined by the criterion of formal association", such as the church or the army. An example from our corpus can be found in example (8):

(8) a. et la peur de **la police impériale**
b. and fear of **the imperial police**

Entities	Occurrences
PER - Mentions	32,338
PER - Chain	3,006
FAC	2,325
TIME	1,836
LOC	1,040
GPE	928
VEH	475
ORG	205
TOTAL	**39,147**

Table 2: The number of occurrences per type of entity.

TIME. This category is absent in the annotations of Bamman et al. (2019). We designed it to annotate temporal information, duration indications, and moments of the day (e.g., *day*, *night*, *morning*). See the example (9) and the example (10) from our corpus:

(9) a. sous **le règne de Louis Xiii**,
b. under **the reign of Louis Xiii**,

(10) a. **Le soir**, il avait mangé un lapin.
b. **At night**, he had eaten a rabbit.

As part of the refinement process, the initial annotations required a thorough revision and cleaning. We had multiple team discussions about many borderline cases, such as whether gods and Greek heroes should be annotated as characters, the status of speaking animals, and the exact distinction between GPE, FAC, and LOC. We meticulously documented every choice made during the annotation process. This documentation is publicly available in an annotation guide[4] and provides a valuable resource for understanding our decisions and methods in characterizing entities in the context of the BookNLP project, based on the initial groundwork provided by the Democrat project. Once the annotation guidelines were finished, the entire corpus was annotated by freshly trained annotators. Their first annotations (comprising 315 tags produced during their training phase) featured an inter-annotator agreement score of 0.38 Cohen's kappa, meaning a fair and almost moderate agreement (Cohen 1960), but showing that this is not a trivial task. With better-trained annotators, values between 0.70 and 0.75 were reached, providing a reasonable basis for further training models. Most of the errors were due to forgotten mentions, and uncertainties about difficult cases (plurals, fuzzy expressions, non-referential entities). Another look at the annotated files by another trained annotator makes a huge difference to get a better and more homogeneous coverage (especially concerning forgotten entities during the initial annotation stage).

After annotation, to facilitate seamless integration with the BookNLP software, the annotations were transformed into a compatible format. We annotated the entity types in TXM (Heiden 2010) because the Democrat corpus is distributed in this format, and later migrated our annotations to brat (Stenetorp et al. 2012), the format used by the team in Berkely. The number of entities in each category can be found in Table 2.

[4] See: `https://github.com/lattice-8094/fr-litbank/blob/main/doc/Manuel_Annotation.pdf`.

3.3 Software Development

LLMs play a prominent role in contemporary NLP. Our implementation of BookNLP-fr is built upon the software from the Multilingual BookNLP Project. Two separate models are developed for the two tasks we perform (namely entity recognition and coreference resolution). Entity recognition is performed before coreference resolution.

Detecting the literary entities, a BiLSTM-CRF model (Bamman et al. 2020; Ju et al. 2018) is fed with contextual embeddings from the CamemBERT model (Martin et al. 2020), which is a BERT-based architecture (Devlin et al. 2019) tailored for French.

For the coreference part, a BiLSTM is also fed with the embeddings from CamemBERT. Following Bamman et al. (2020), who in turn follows Lee et al. (2017), the BiLSTM architecture is attached to a feedforward network in which the probability that two mentions (detected entities) are coreferent with each other is evaluated. Mentions are linked to their highest scoring antecedent (a null-antecedent is always an option), and coreference chains are defined as the transitive closures of links.

For each model, we split the corpus into training (80%), development (10%), and test (10%) sets. The results can be found in section 4.

While event annotation remains a focal point, challenges persist, primarily due to limitations in performance and the inherently ambiguous nature of defining events. The elusive nature of the concept makes it challenging to generate consistently relevant and usable results. As for quotation identification, we acknowledge the need to integrate speaker recognition for a more comprehensive understanding of textual nuances.

Given these considerations, we have more specifically directed our efforts toward optimizing modules for entity recognition and coreference resolution. This focus allows us to refine and train models that are specifically accurate in identifying and linking entities within a given text, contributing to the effectiveness of BookNLP-fr for downstream tasks (like subgenre classification, see section 5).

4. Results and Evaluation

In this section, we present the results of our BookNLP-fr modules for entity recognition and coreference resolution on literary texts.

4.1 Named Entity Recognition Evaluation

Table 3 reports our results for entity recognition, traditionally measured through precision (the percentage of entities correctly recognized among those recognized) and recall (the percentage of entities correctly recognized among those to be recognized). Please note that ORG is absent from this evaluation, because due to an uneven distribution of this tag in different texts, it was only present seven times in the test corpus, making the estimation of precision and recall unreliable.

When assessing the model's performance, a higher precision relative to recall suggests that the model is more likely to make accurate predictions when identifying literary entities. Precision denotes the percentage of correctly predicted literary entities among

	Precision	Recall	F_1
PER	85.0	92.1	88.4
LOC	59.4	54.3	56.8
FAC	73.4	66.0	69.5
TIME	75.3	36.4	49.1
VEH	68.9	63.6	66,1
GPE	68.2	52.9	59,6

Table 3: Entity recognition evaluation of BookNLP-fr on literary texts.

	BookNLP-fr			CamemBERT-NER		
POS Tag	Precision	Recall	F_1	Precision	Recall	F_1
PROP	82.5	79.2	**80.8**	91.85	72.05	80.75
NOM	74.9	74.7	**74.8**	96.32	14.17	24.70
PRON	86.3	89.5	**87.9**	100.00	0.10	0.20
ALL	82.39	83.88	**83.13**	92.58	7.92	14.59

Table 4: Comparison on litbank-fr for PER recognition performance between BookNLP-fr and CamemBERT-NER.

all entities predicted by the model. High precision is advantageous, ensuring that the identified literary entities are more likely to be accurate, albeit at the potential cost of missing some relevant entities (lower recall). Prioritizing precision in this context aids in minimizing false positives, thereby enhancing the reliability of the identified literary entities. It is important to highlight that literary entities differ from typical named entities in NLP, displaying a much larger range of possibilities. Consequently, the obtained results, though seemingly divergent from NLP standards, represent a pioneering achievement in the analysis of French fiction, as this is the first study of its kind. Some scores may appear modest compared to the state-of-the-art, particularly regarding the recall for TIME expressions. This is due to the extensive diversity of time expressions in our corpus, which is far more varied than in the traditional news corpora typically used in NLP, coupled with the limited number of examples in the training corpus (see Table 4 for a comparison with a state-of-the-art system). Nevertheless, we have opted to report these scores for the sake of comprehensiveness. In the near future, we will strive to expand the coverage of our system, aiming to achieve improved recall across various categories beyond PER.

As a baseline, we ran the CamemBERT-NER model[5], which is a NER model fine-tuned from CamemBERT on the wikiner-fr dataset. Table 4 shows the baseline performance compared to BookNLP-fr. The results show that BookNLP-fr is as good as the fine-tuned model for proper name recognition, but it captures much more by including pronouns and common nouns, which the baseline model does not handle at all. The F_1 score for the detection of PROP/NOM/PRON mentions reaches 83.13, which is in line with the English BookNLP (88.3). BookNLP-fr thus demonstrates its robustness for the classic task of proper name recognition. However, the real value of our model lies in its ability to go beyond this to capture the full spectrum of what constitutes a character in novels. This aligns with Woloch (2003)'s concept of the character space as "the encounter between an individual human personality and a determined space and

[5] See: https://huggingface.co/Jean-Baptiste/camembert-ner.

Metrics	F_1
MUC	88,0
B^3	69,2
$CEAF_e$	71.8

Table 5: Coreference resolution evaluation of BookNLP-fr on literary texts with an average F_1 score of 76.4, calculated as the mean of the three metrics.

position within the narrative as a whole", allowing automatic detection and analysis of the distribution of character mentions throughout the narrative (Barré et al. 2023).

4.2 Coreference Resolution Evaluation

Table 5 presents the evaluation metrics for coreference resolution using BookNLP-fr on our test corpus. Three key metrics, namely MUC, B^3, and $CEAF_e$, are employed to assess its performance. As coreference chains are complex to model, different evaluation metrics are necessary to get a global image of the system's performance. We refer to Luo and Pradhan (2016) for a comprehensible explanation of these metrics. Our average F_1 score is 76.4, calculated as the mean of the three metrics. The reported scores suggest a commendable performance, but the practical utility in the context of literary analysis should be further explored based on the specific goals of the research or application. Note that the English BookNLP yields 79.3 in performance for the same task.

The challenge of duplication arises when the model detects the same character multiple times within the analyzed text. In some instances, among the top five literary entities identified by the model, there may be cases where two or more main characters share the same name or attributes. While this duplication might initially raise concerns, e.g. in the study of character networks (Perri et al. 2022) or the overall number of characters in novels, it may not pose a significant issue when the focus is on character characterization. For example, in studies of the representation of male and female characters, the output of BookNLP has proven useful (e.g., Gong et al. 2022; Hudspeth et al. 2024; Naguib et al. 2022; Toro Isaza et al. 2023; Underwood et al. 2018; van Zundert et al. 2023; Vianne et al. 2023).

Also in the following case study, the primary objective is not to pinpoint unique and distinct characters, but rather to establish a proxy for characterization as a whole. Our goal is to capture the prevalence and significance of certain characters across various texts and literary works. Hence, the emphasis lies more on character representation and the overall impact of these characters on the literary landscape, rather than on identifying entirely separate and non-repeating characters.

5. Case Study: Genre Classification Using BookNLP-fr Features

5.1 Introduction

This case study aims to demonstrate that BookNLP-fr can be of significant assistance in the realm of computational literary studies (CLS). We illustrate this assertion through a canonical issue in CLS: the automatic detection of literary genres. Historically, the division of novels into specific subgenres has been a classification practice employed by literary stakeholders such as librarians, editors, and critics. This practice is partly justified by a specific textual component that relates to the spatiotemporal framework, characters, themes, or narrative progression.

Genre is a central concept in poetics, successively defined by theorists from Aristotle to the structuralists through the romantics and Russian formalists (i.a. Aristote 1990; Bachtin [1987] 2006; Genette 1986; Schlegel et al. 1996). From our computational standpoint, structuralists have offered intriguing definitions. For example, Schaeffer (1989, 73) defines genericity as an "internalized norm that motivates the transition from a class of texts to an individual text conforming to certain traits of that class". There could be a set of textual procedures internal to works, and the mission of CLS would be to find the best ways to account for that fact. However, the norms or formal rules of subgenres cannot be solely boiled down to formal or thematic rules. The sociological approach, as illustrated by Bourdieu (1979), emphasizes the influence of reader communities in defining genres, examining power dynamics and the accompanying aesthetic hierarchies in the literary field. Nevertheless, these norms do indeed exist, as they enable a work to conform to the established and shared use of a "horizon of expectations" (Jauß 1982, 22) audience, which might induce authors to adhere to certain expected norms and styles.

Various studies have devised strategies to automatically identify subgenres. Selected studies have employed methods such as the bag-of-words (BoW) (Hettinger et al. 2016; Underwood 2019) or topic modeling (Schöch 2017; van Zundert et al. 2022) to find subgenre similarities between texts. In addition to these basic features, researchers utilize machine learning techniques in a supervised setting, employing methods such as logistic regression or support vector machines when ground truth is available. However, the challenge often arises from the potential incompleteness or temporal bias of these ground truths. Unsupervised learning approaches and clustering methods have also enabled the exploration of hybrid texts that belong to multiple subgenres, as demonstrated by studies like (Calvo Tello 2021; Sobchuk and Šeļa 2024). In our case study, we will rely on a corpus with predefined labels, while acknowledging the idea that subgenres are not monolithic categories. Thus, the objective is not so much to demonstrate the validity of subgenre labels, which are often incomplete or limiting in reality, but rather to show that the interpretability of errors in automatic classification can lead us to a more nuanced and comprehensive understanding of the subgenre phenomenon.

Despite recent advancements in NLP, the bag-of-words approach remains largely unchanged. This is because many tools, including document embeddings, are not easily interpretable and are optimized for short texts. In this context, we present in the next section a method that aims to find a balance between the use of state-of-the-art methods for literary text processing and their interpretability.

5.2 Method

5.2.1 Corpus and Subgenre Labels

Our case study is built on one of the largest corpora for fiction in French: the corpus *Chapitres*, a corpus of nearly 3,000 French novels (Leblond 2022). The investigated period

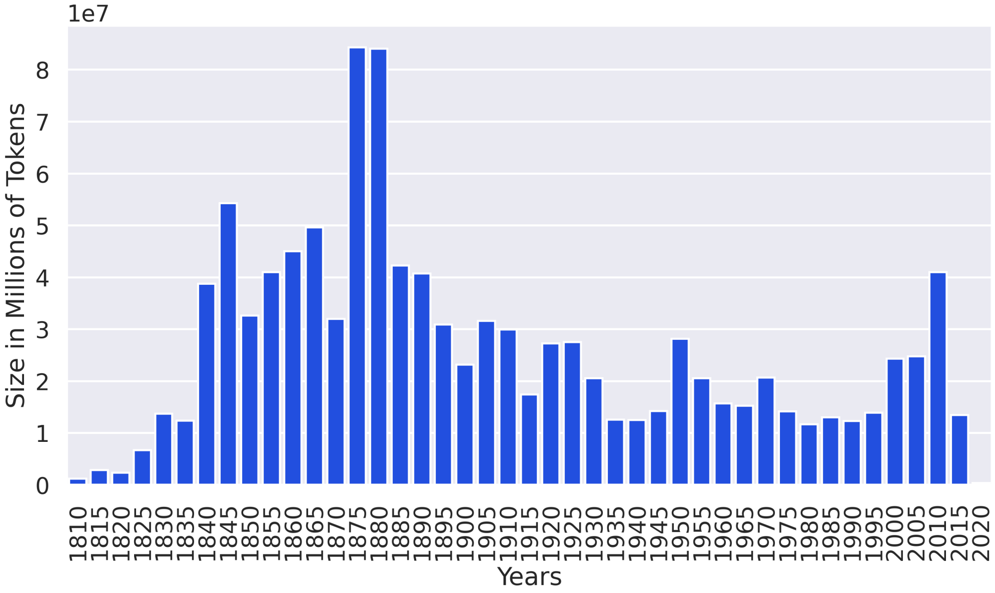

Figure 1: Distribution of the number of tokens over time.

covers over two centuries of novel production, from the 19th to the 20th century, as can be seen in Figure 1.

Approximately two-thirds of Chapitres is annotated with subgenre labels. This annotation is based on the classification of the French National Library (BNF). We choose to concentrate our analysis on the five most prevalent subgenres within the corpus: adventure novels, romance, detective fiction, youth literature, and memoirs. The validity of these labels is not clearly established, as the practices of the BNF for assigning these labels have not been systematized nor standardized. Therefore, there is no ground truth *per se*, but our supervised approach described in subsubsection 5.2.3 aims precisely to understand the boundaries of subgenres.

5.2.2 Textual Features

The BoW method stands out as the default feature extraction technique, as it allows scholars to have an easy task to implement without requiring intensive computational resources (GPU, RAM). Underwood (2019) demonstrated that the BoW approach was highly effective in classifying subgenres such as Gothic fiction, detective stories, and even science fiction.

Nevertheless, although this method proves valuable in specific contexts, it is not without two limitations. First, it does not consider the word order within the text. This limitation means that the sequential arrangement of words, which is crucial for capturing the nuances of literary elements like plot and narrative structure, is ignored. Second, there is a risk of overfitting to the idiolects of writers, particularly when emphasizing the most frequent words (MFW). Additionally, these tools may inadvertently capture chronolectal aspects, as it is established that the approximate writing date of a book can be predicted based on the prevalence of certain most frequent words (Seminck et al. 2022).

In this paper, we rely on two distinct feature extraction approaches: the classic BoW as a control experiment and the BookNLP-fr one, which we will implement as follows.

The idea is based on a previous study by Kohlmeyer et al. (2021), where researchers demonstrated the limitations of traditional document embeddings (optimized for shorter texts) in capturing the complex facets of novels (such as time, place, atmosphere, style, and plot). To address this problem, they propose to use multiple embeddings reflecting different facets, splitting the text semantically rather than sequentially. Inspired by these findings, we adapted their method to evaluate the impact of these features on subgenre classification when contrasted with the traditional BoW approach.

The method runs the BookNLP-fr pipeline on the corpus texts. On the one hand, it allows us to automatically retrieve information related to space-time, notably with the set of LOC, FAC, GPE, TIME, and VEH. On the other hand, it provides information related to characterization, including all verbs for which characters are patients (PATIENT) or agents (AGENT), as well as the set of adjectives that will characterize them (ADJ). Thus, two types of features are under consideration:

- For the BoW, we relied on the 600 most frequent lemmas, excluding the first 200, which comprise non-informative stop words not relevant to our subgenre case study. They could have been relevant if we wanted to acknowledge the authors who wrote in a specific subgenre, but it is not our goal here, and we will discuss how we handled this bias in subsubsection 5.2.3.
- For the BookNLP-fr features, we compiled lists of words extracted by BookNLP-fr for each novel. We then obtained vector representations using a Paragraph Vectors model (Le and Mikolov 2014) (Doc2Vec) trained on a subset of our novel dataset. Two vector embeddings of 300 dimensions were generated: one for characterization (AGENT, PATIENT, ADJ) and one for space and time (LOC, FAC, GPE, TIME, VEH).

Therefore, we obtained two datasets for training, one with 600 dimensions representing the 600 most frequent lemmas, and the other with also 600 dimensions representing the two concatenated Doc2Vec vectors, one for the characterization and one for the space and time.

5.2.3 Modeling

We opted for a Support Vector Machine (SVM) as it has been demonstrated that these models obtain the best performance in classifying literary texts (Yu 2008), and more specifically literary subgenres (Hettinger et al. 2016). In this paper, we used the implementation of Pedregosa et al. (2011). The SVM doesn't perform multiclassification *per se*, but it classifies each subgenre against the others in a binary classification and then aggregates the results. Therefore, we do not have a single classification, but rather:

$$\frac{n_{\text{classes}} \cdot (n_{\text{classes}} - 1)}{2}$$

With our five subgenres, this implementation results in ten different classifications.

Considering our task of subgenre classification, we wanted to limit idiolectal bias, especially for the model trained on the BoW. To do so, we implemented Scikit-learn's group strategy. All works by the same author (group) were placed in the same fold. Thus, each group appears exactly once in the test set across all folds. Since SVM models

	Precision	Recall	F_1	Support	Accuracy
Children	0.75	0.75	0.75	130	
Memoirs	0.79	0.82	0.80	130	
Detective	0.67	0.68	0.67	130	
Adventure	0.60	0.65	0.62	130	
Romance	0.84	0.72	0.80	130	
Full Dataset				650	**0.72**

Table 6: Classification report for BoW.

	Precision	Recall	F_1	Support	Accuracy
Children	0.65	0.79	0.71	130	
Memoirs	0.78	0.89	0.84	130	
Detective	0.68	0.70	0.70	130	
Adventure	0.73	0.73	0.73	130	
Romance	0.90	0.65	0.75	130	
Full Dataset				650	**0.75**

Table 7: Classification report for BookNLP-fr features.

are quite sensitive working with imbalanced classes, we re-balanced the classes before implementing the classification by randomly taking 130 novels for each subgenre. We implemented this selection one hundred times, and for each resulting sample, the model was run in a 5-fold cross-validation setting. The following results are aggregated from this process.

5.3 Results

5.3.1 BoW vs. BookNLP-fr Features

Table 6 and Table 7 display the classification report of the models' evaluation on the test set. Both models achieve good results: 72% accuracy for the BoW-based model and 75% for the BookNLP-fr-based model. This means that our models are capable to correctly identify the subgenre three out of four times, whereas a random baseline yields an accuracy score of 0.2. The main result here is that differences exist among our subgenres, whether from the perspective of text structure with MFW or from a semantic standpoint with BookNLP-fr. The fact that the BookNLP-fr-based model obtains an additional 3 points of accuracy might not be revolutionary, but the primary argument for this type of feature extraction lies more in the interpretation of features, as discussed in subsection 5.4.

To enhance our comprehension of how the models behave and the nature of their errors, we visualize their confusion matrices in Figure 2 and Figure 3. The x-axis represents the predicted subgenre, while the y-axis represents the expected subgenre. A perfect classification would display a diagonal filled with 130 correct predictions for each subgenre.

We observe that both models have quite similar error patterns, and one distinct scenario stands out: Both models predict 'Adventure' instead of 'Detective' (23 errors for BoW, 21 for BookNLP-fr). These common errors are quite understandable, since these two

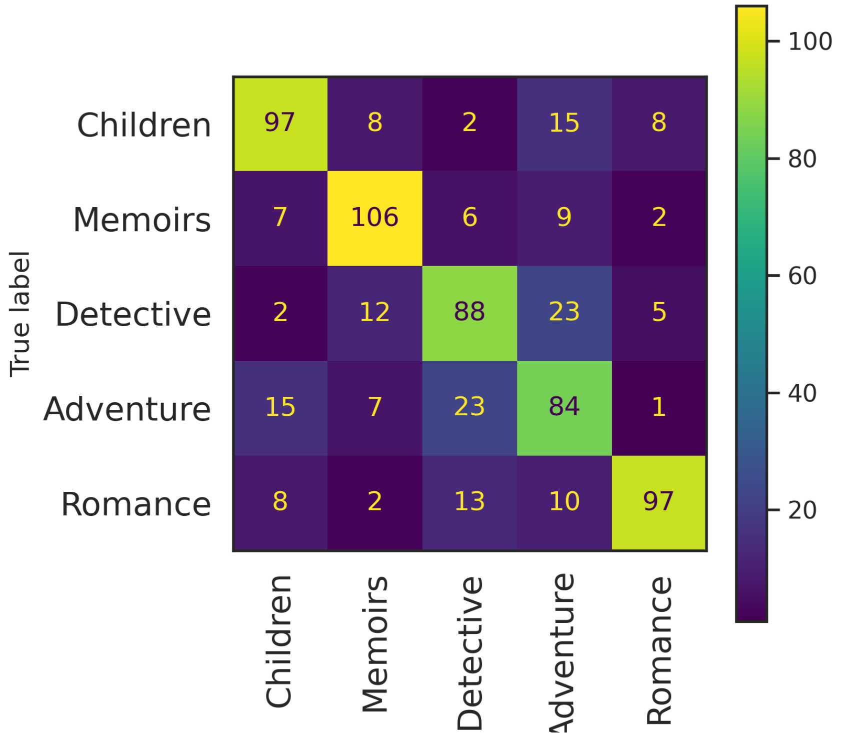

Figure 2: Confusion matrix for BoW.

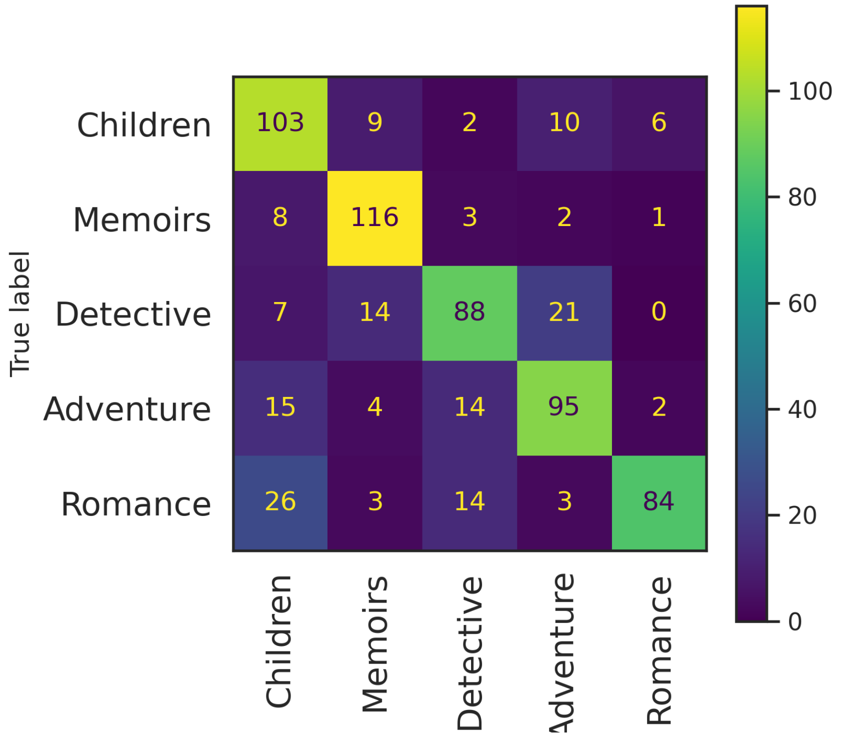

Figure 3: Confusion matrix for BookNLP-fr features.

BookNLP-fr Features	Accuracy
LOC	0.45
FAC	0.59
VEH	0.42
GPE	0.47
TIME	0.50
PATIENT	0.52
AGENT	0.62
ADJ	0.50
Baseline	0.2

Table 8: BookNLP-fr features accuracy.

subgenres share many similarities, including a penchant for suspense and violent action, which could confuse the models.

Another scenario seemed highly instructive for analysis: The errors made by the models when predicting the label 'Children', but the expected subgenre is 'Romance'. The BoW model performs quite well with 8 errors, but the BookNLP-fr-based model makes 26 errors. Thus, the semantic model faces more challenges in distinguishing between these two subgenres, which makes sense as both subgenres are characterized by themes centered around emotions and relationships between characters, which are common features of both subgenres.

5.3.2 BookNLP-fr Features Accuracy for Subgenre Classification

The objective of this section is to evaluate, whether specific individual features from BookNLP-fr can classify our subgenres, and to attempt to interpret the differences in performance for each. Here, each pipeline is trained with a Doc2Vec vector of 300 dimensions for each feature type.

A first obvious observation is that all of our models perform at least twice as well as the baseline. The information contained in each of these features is therefore highly relevant from a subgenre perspective. The 'VEH' class lags a bit behind (accuracy of 0.42), which may suggest that vehicles are not decisively discriminating among our subgenres. However, it is our least represented class in our texts, and therefore, there may not be enough data. Very good results are obtained for 'FAC' (0.59) and 'AGENT' (0.62). This indicates that subgenres distinguish well in terms of mentioned buildings or verbs where the character is agentive, meaning that the type of action a character takes is specific to each subgenre.

Interestingly, the misclassifications shows the same pattern (misclassification of 'Adventure' instead of 'Detective' and 'Children' instead of 'Romance', see the confusion matrices in the Appendix A for each individual feature), but the error rates vary depending on the features used. This can provide a lot of information about the differences and similarities between certain subgenres. The subsection 5.4 offers an interpretation that closely examines these anomalies.

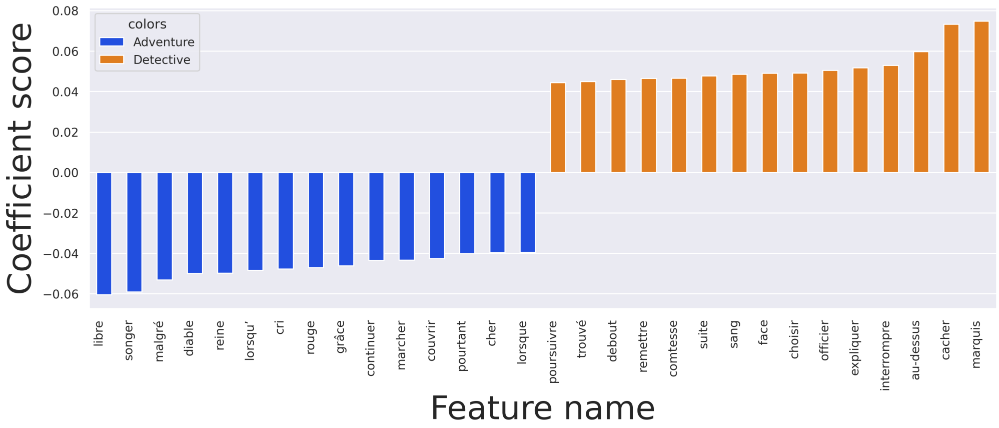

Figure 4: BoW discriminant features for ‘Adventure’ vs. ‘Detective’ classification.

5.4 Interpretability

This section explores the interpretation of the two SVM models (BoW-based and BookNLP-fr-based). It focuses on the misclassification of ‘Adventure’ instead of ‘Detective’.

One of the advantages of the SVM pipeline is the ability to investigate the statistical inference of the models when the kernel is in linear mode. The SVM searches for the plane in the latent space of words that best separates our two categories. Each dimension receives a coefficient, with a negative sign if the coefficient is used to predict a specific class, and a positive sign for the other. For the BoW-based model, this is quite straightforward, as a coefficient is assigned to each word (see Figure 4).

Looking at the coefficients assigned for the ‘Adventure’ vs. ‘Detective’ classification, we find some relevant elements, such as the presence of the word ‘free’ (‘libre’) as the most discriminant word for assigning the ‘Adventure’ label. Apart from that, with the possible exception of the noun ‘cry’ (‘cri’), which could signify adventure, few clues remain. Verbs such as ‘dream’ (‘rêver’), ‘walk’ (‘balader’), ‘continue’ (‘continuer’), or conjunctions like ‘when’ (‘lorsque’), ‘despite’ (‘malgré’), and ‘yet’ (‘pourtant’) are not really characteristic of adventure novels. It is difficult to draw conclusions, except that these less significant coefficients seem to indicate the model’s difficulty in distinguishing between the two subgenres.

For the BookNLP-fr-based model, it is a bit more complex since the coefficients are assigned to each dimension of the Doc2Vec vectors. Therefore, we aggregated the coefficients by feature type to gain a more concrete overview of the results. Figure 5 illustrates the sum of all coefficients for each feature extracted by BookNLP-fr. We conducted a t-test to confirm that the difference between the means of the populations is statistically significant. Taking adjectives as an example (t-statistic: 28.7; p-value: 2.25×10^{-180}), we observe that the model relies more on these dimensions to assign the label ‘Detective’ compared to ‘Adventure’.

This could be explained by the strong emphasis placed on character psychology in detective novels, especially those involving criminals and detectives. For instance, in *Maigret et le tueur* (1969), George Simenon’s beloved detective (*Maigret*) is frequently

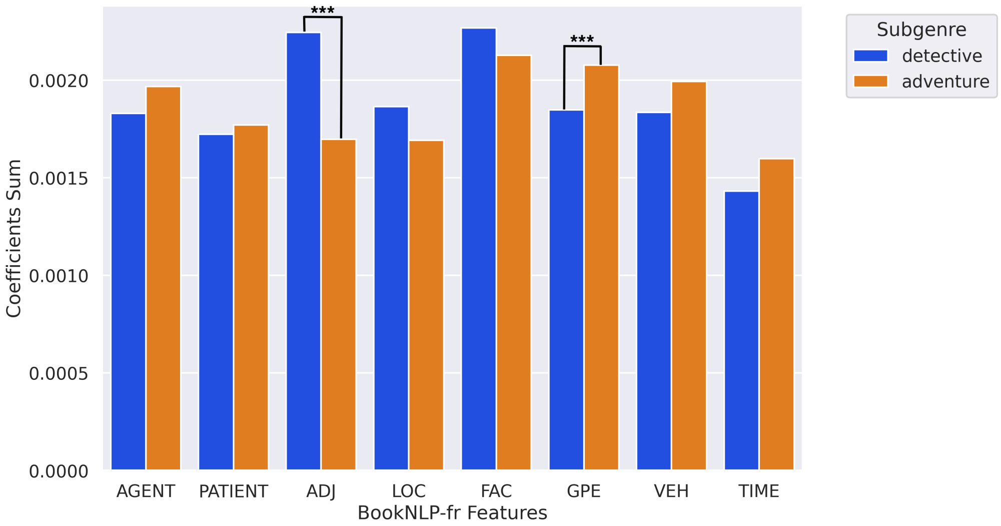

Figure 5: BookNLP-fr discriminant features for 'Adventure' vs. 'Detective' classification. '***' meaning $p < 0.001$.

characterized as 'wise' ('sage'), 'whimsical' ('fantaisiste'), or even 'happy' ('heureux'), while criminals are 'suspicious' ('suspect') or 'villainous' ('méchant'). This does not imply a lack of characterization in adventure novels but rather suggests that it is not a distinctive feature of the subgenre compared to detective novels.

Considering GPEs (t-statistic: -21.0; p-value: 8.49×10^{-98}), the reasoning is inverse: The model relies slightly more on the dimensions of the GPE vector to assign the 'Adventure' label than the 'Detective' label. This makes sense when examining GPEs for example in *Les trappeurs de l'Arkansas* by Gustave Aimard (1857): 'Hermosillo', 'America', 'the New World', 'Guadalajara', 'Mexico', etc. The novel heavily emphasizes exotic locations and mentions places in the American or Mexican West for this purpose. GPEs in detective novels are more commonplace, as these novels often take place in France, typically in an urban setting.

Thus, the model has learned that certain dimensions of characterization are more strongly associated with a particular subgenre (such as adjectives for detective novels), and that certain dimensions of the GPE or TIME vector are important for assigning the 'Adventure' label. In the following we generalize our approach to the entire classification process.

Examining the behavior of the coefficients when aggregated for the ten classifications, we can observe the graph shown in Figure 6. This graph depicts the model coefficients after training based on the vectors of each facet, using a dataset of 2,400 dimensions. We consider this graph as a dive into the model's inferences, where it will assign more weight to certain categories to assign a specific subgenre.

For example, it is observed that the value of 'FAC' is very high for the detective genre, indicating a particular specificity for this subgenre. Details of locations, crime scenes, investigations in specific places, detective offices, interrogation rooms, etc. are distinguishing elements of this subgenre. The same applies to 'GPE' for the 'Adventure' label, as seen previously, with an emphasis on exoticism that may play a role here, even though 'LOC' and 'FAC' do not show significant differentiation from this perspective.

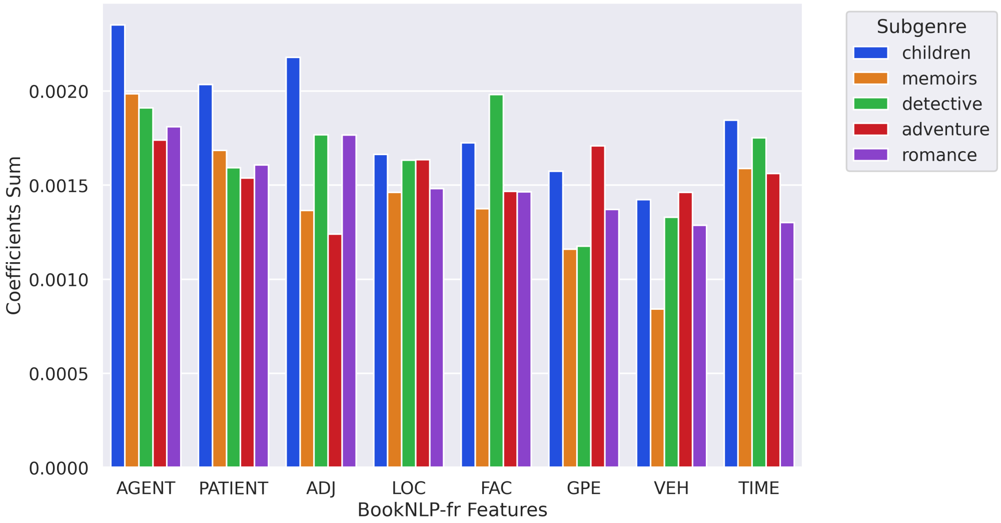

Figure 6: BookNLP-fr discriminant features for the classification.

Conversely, for 'Romance' and the 'TIME' vector, where the coefficients for these vectors lag behind other subgenres. Examples of time in romance novels may be used more to describe emotional moments or stages in relationships rather than to highlight complex temporal events. Consequently, the model might perceive that the 'TIME' vector is less discriminative for this category.

We have thus demonstrated that the BoW-based classification approach is challenging to interpret, as certain highly discriminating words do not appear to bring about key distinctions between the subgenres. The BookNLP-fr-based method may offer an insightful understanding of the specificities that differentiate one subgenre from another. Both approaches do not completely substitute each other, since we are examining features of different nature (vocabulary vs. semantic), but they can complement each other to enhance interpretability.

Diving into the model's indications, several types of features were observed to interpret the model's inferences. Many differences among the features were noticed, although we did not have the space to interpret all of them in this article. Much work remains to be done, and new experiments should be considered, for instance going beyond the SVM, including the use of deep neural networks and textual deconvolution saliency Vanni et al. (2018), which could facilitate the return to close reading based on the embeddings derived from BookNLP-fr data.

6. Discussion

6.1 Working with Imperfect Annotations

The utilization of computers for annotating literary texts has profoundly changed the landscape of literary studies, enabling the annotation of vast amounts of texts with unprecedented efficiency. This enables the community to address research questions that were out of reach before, such as a study at scale of characters with disabilities (Dubnicek et al. 2018), or the quantitative analysis of characters in fanfiction (Milli and Bamman 2016), and a quantitative, diachronic study of things appearing in fiction (Piper and

Bagga 2022). However, this advancement is not without its challenges, particularly in the context of the inherent errors that may accompany automated annotation processes. This poses a twofold challenge for researchers engaged in the field of CLS.

Firstly, ensuring the reliability of studies based on imperfect annotations is a critical concern. Scholars must grapple with the task of guaranteeing that errors, though present, remain at a marginal level and do not compromise the validity of their research findings. This necessitates a careful balance between the benefits of computational efficiency and the maintenance of accuracy in annotations. Researchers are challenged to develop methods and quality control measures that safeguard against the potential pitfalls introduced by errors in the annotation process.

Secondly, the acceptance of computational approaches by literary scholars is not guaranteed, as the traditional paradigm within literary studies often revolves around meticulous, supposedly perfect annotations. The shift to working with non-perfect annotations, even if the errors are marginal, represents a departure from the established norm. This cultural shift within the academic community poses a psychological barrier, as literary scholars may be hesitant to fully embrace computational methods if they perceive a compromise in the level of precision to which they are accustomed.

Addressing these challenges requires not only the refinement of computational tools for annotation but also a broader cultural shift within the academic community. There is a need for transparent communication about the limitations of automated annotation processes, the establishment of best practices for mitigating errors, and the development of strategies to ensure that computational approaches align with the standards expected both in literary studies and in computer science.

6.2 Maintaining Annotation Tools in the Era of LLMs

The field of CLS is currently grappling with a significant challenge due to the rapid evolution of NLP, particularly with the proliferation of LLMs. The continuous emergence of new LLMs has led to an accelerated pace of research in the domain. While this dynamism brings about positive outcomes, such as increased research activity, the introduction of novel tasks, and the generation of new results, it also presents several inherent dangers.

One primary challenge lies in the technical aspect of keeping annotation tools up to date amidst the constant production of new LLMs by the research community and the industry. There is a delicate balance to strike, ensuring that annotation systems remain up-to-date, without expending an excessive amount of resources on incessantly adapting to the latest trends in LLM development. The challenge here is not just about technological compatibility but also about efficiently managing the resources required for frequent updates and integrations, and to produce software that is usable by a large community (i.e., software should not be dependent on an unreasonably heavy computer infrastructure).

A more critical concern revolves around the need to guarantee the reproducibility of research outcomes. The rapid evolution of LLMs implies that a specific version in use today may become obsolete or unavailable tomorrow. This raises the risk that crucial details, such as the corpus utilized, configuration parameters, and hyperparameters of the model, may not be adequately documented in research reports. Ensuring reproducibility becomes a substantial challenge as the landscape of LLMs continues to evolve, necessitating a concerted effort to establish standardized practices for reporting model specifications and associated details.

In addressing these challenges, we believe it is crucial to focus not only on technical aspects but also on developing robust frameworks for documentation and reproducibility. Establishing clear guidelines for reporting model specifications, documenting corpus details, and archiving relevant information becomes paramount for the field.

7. Conclusion

In this paper, we introduced the BookNLP-fr pipeline, with a particular emphasis on entity recognition and coreference resolution. Demonstrating its practical utility, we illustrated how this software facilitates the analysis of extensive French literary corpora, relying on semantic features unique to the texts under examination. Through this study, we hope to show the potential of NLP in analyzing large literary corpora, to go beyond purely statistical approaches and to overcome bias by taking into account an unprecedented number of texts and not only the reduced set of texts of the literary canon. In concrete terms, we distinguish three research directions, all of which carry the above-described desire for large-scale generalization:

1. Studies on the characteristics of literary genre: BookNLP-fr can be used to retrieve textual features of a semantic nature, in particular entities that provide information on the spatio-temporal setting of the story. The latter are very important for determining literary genres. For example, adventure novels have a very specific spatio-temporal setting (the emphasis is on the importance of geographical disorientation), while romance novels take place in a more urban, modern setting. The BookNLP-fr tools could thus be crucial for automatic classification.

2. Characterization: Coreference chains with character mentions allow us to recover how each character is portrayed. In this way, we can study the differences between certain types of characters on a large scale. For example, it is possible to report on how men and women have been characterized in literature over time (e.g., Naguib et al. 2022; Vianne et al. 2023) or what role secondary characters actually play in the narrative (Barré et al. 2023). To cite other examples: A tool like BookNLP makes it possible to study how characters with disabilities are presented (Dubnicek et al. 2018) or to carry out a quantitative analysis of characters in fanfiction (Milli and Bamman 2016).

3. Detection of specific scenes: BookNLP could be capable of detecting specific scenes in novels; these could be defined by one or more characters gravitating around a precise location and carrying out particular actions. This scene detection, understood as a small narrative unit, could enable us to better understand the workings of the plot by breaking down its layout over the course of the story.

Future work on the BookNLP-fr pipeline will include a renewed exploration of the concepts of events and scenes, aiming to establish an annotation framework that aligns with literary perspectives. Additionally, we plan to address the question of quotation analysis and attribution. Finally, a key focus will be to ensure that results undergo scientific evaluation and that recent advancements in NLP can be continuously integrated, while preserving the distinctive nature of literary works and literary studies. In that way, BookNLP-fr can play a significant role in the domains of automatic literary analysis and cultural analysis. Literary questions, one even more exciting and ambitious than the other, can finally be addressed automatically on a large scale.

8. Data Availability

Data used for the research has been archived and is persistently available at: https://doi.org/10.5281/zenodo.14018430 for the *French LitBank Corpus* and https://doi.org/10.5281/zenodo.7446727 for the *Corpus Chapitres*.

9. Software Availability

All code created and used in this research has been published at: https://github.com/lattice-8094/DEV_BOOKNLP_FR/tree/v0.1.0 and https://github.com/crazyjeannot/jcls_booknlp_subgenres for the case study. It has been archived and is persistently available at: https://doi.org/10.5281/zenodo.14018556 and https://doi.org/10.5281/zenodo.14135475 for the case study.

10. Acknowledgements

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA0001 (PRAIRIE 3IA Institute).

11. Author Contributions

Frédérique Mélanie-Becquet: Conceptualization, Data curation, Supervision

Jean Barré: Formal analysis, Writing – review & editing

Olga Seminck: Formal analysis, Writing – review & editing

Clément Plancq: Conceptualization, Software

Marco Naguib: Conceptualization, Software

Martial Pastor: Conceptualization, Software

Thierry Poibeau: Conceptualization, Writing – original draft, Writing – review & editing, Supervision

References

Aristote (1990). *Poétique*. Le Livre de Poche. Librairie Générale Française.

Bachtin, Michail Michajlovič [1987] (2006). *Esthétique et théorie du roman*. Gallimard.

Bamman, David (2020). *Multilingual BookNLP: Building a Literary NLP Pipeline across Languages*. https://apps.neh.gov/publicquery/main.aspx?f=1&gn=HAA-271654-20 (visited on 01/17/2024).

— (2021). *BookNLP*. https://github.com/booknlp/booknlp (visited on 11/04/2024).

Bamman, David, Olivia Lewke, and Anya Mansoor (2020). “An Annotated Dataset of Coreference in English Literature”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, 44–54. https://aclanthology.org/2020.lrec-1.6 (visited on 10/13/2024).

Bamman, David, 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*, 2138–2144. 10.18653/v1/N19-1220.

Bamman, David, 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*, 370–379. 10.3115/v1/P14-1035.

Barré, Jean, Pedro Cabrera Ramírez, Frédérique Mélanie, and Ioanna Galleron (2023). “Pour une détection automatique de l’espace textuel des personnages romanesques”. In: *Humanistica 2023*, 56–61. https://hal.science/hal-04105537 (visited on 10/13/2024).

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). “quanteda: An R Package for the Quantitative Analysis of Textual Data”. In: *Journal of Open Source Software* 3 (30), 774–774. 10.21105/joss.00774.

Bird, Steven, Ewan Klein, and Edward Loper (2019). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*. https://www.nltk.org/book/ch00.html (visited on 10/13/2024).

Bourdieu, Pierre (1979). *La distinction: critique sociale du jugement*. Le Sens commun 58. Éditions de Minuit.

Brunner, Annelen, Tanja Tu, Lukas Weimer, and Fotis Jannidis (2020). “To BERT or Not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of Four Types of Speech, Thought and Writing Representation”. In: *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. https://ceur-ws.org/Vol-2624/paper5.pdf (visited on 11/04/2024).

Calvo Tello, José (2021). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press. 10.1515/9783839459256.

Cohen, Jacob (1960). “A Coefficient of Agreement for Nominal Scales”. In: *Educational and Psychological Measurement* 20 (1), 37–46.

Dekker, Niels, Tobias Kuhn, and Marieke van Erp (2019). “Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels”. In: *PeerJ Computer Science* 5, e189. 10.7717/peerj-cs.189.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. 10.18653/v1/N19-1423.

Dubnicek, Ryan, Ted Underwood, and J. Stephen Downie (2018). "Creating A Disability Corpus for Literary Analysis: Pilot Classification Experiments". In: *iConference 2018 Proceedings*. https://core.ac.uk/download/pdf/159610829.pdf (visited on 10/13/2024).

Durandard, Noé, Viet Anh Tran, Gaspard Michel, and Elena Epure (2023). "Automatic Annotation of Direct Speech in Written French Narratives". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 7129–7147. 10.18653/v1/2023.acl-long.393.

Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). "Stylometry with R: A package for Computational Text Analysis". In: *R Journal* 8 (1), 107–121. 10.32614/RJ-2016-007.

Emelyanov, Anton and Ekterina Artemova (2019). "Multilingual Named Entity Recognition Using Pretrained Embeddings, Attention Mechanism and NCRF". In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 94–99. 10.18653/v1/W19-3713.

Genette, Gérard (1986). "Introduction à l'architexte". In: *Théorie des genres*. Ed. by Gérard Genette and Tzvetan Todorov. Points 181. Édition du Seuil.

Gong, Xiaoyun, Yuxi Lin, Ye Ding, and Lauren Klein (2022). "Gender and Power in Japanese Light Novels". In: *Proceedings of Computational Humanities Research*, 256–267. https://ceur-ws.org/Vol-3290/short_paper1101.pdf (visited on 10/13/2024).

Heiden, Serge (2010). "The TXM Platform: Building Open-source Textual Analysis Software Compatible with the TEI Encoding Scheme". In: *24th Pacific Asia Conference on Language, Information and Computation*. 2, 389–398. https://aclanthology.org/Y10-1044 (visited on 10/13/2024).

Hettinger, Lena, Fotis Jannidis, Isabella Reger, and Andreas Hotho (2016). "Significance Testing for the Classification of Literary Subgenres". In: *Book of Abstracts of DH 2016*. https://dh-abstracts.library.cmu.edu/works/2630 (visited on 01/18/2024).

Hogenboom, Frederik, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron (2016). "A Survey of Event Extraction Methods from Text for Decision Support Systems". In: *Decision Support Systems* 85, 12–22. 10.1016/j.dss.2016.02.006.

Hudspeth, Marisa, Sam Kovaly, Minhwa Lee, Chau Pham, and Przemyslaw Grabowicz (2024). *Gender and Power in Latin Narratives*. https://www.marisahudspeth.com/_files/ugd/8e4ffa_8720b9ffa7694f76adedf5ad9e07248d.pdf (visited on 10/13/2024).

Jauß, Hans Robert (1982). *Toward an Aesthetic of Reception*. Trans. by Timothy Bahti. University of Minnesota Press.

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (2019). "BERT for Coreference Resolution: Baselines and Analysis". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5803–5808. 10.18653/v1/D19-1588.

Ju, Meizhi, Makoto Miwa, and Sophia Ananiadou (2018). "A Neural Layered Model for Nested Named Entity Recognition". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1446–1459. 10.18653/v1/N18-1131.

Kohlmeyer, Lasse, Tim Repke, and Ralf Krestel (2021). "Novel Views on Novels: Embedding Multiple Facets of Long Texts". In: *2021 Association for Computing Machinery.* https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2021_kohlmeyer_novel.pdf (visited on 10/13/2024).

Landragin, Frédéric (2016). "Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT)". In: *Bulletin de l'Association Française pour l'Intelligence Artificielle* 92, 11–15. https://www.researchgate.net/publication/306111513_Description_modelisation_et_detection_automatique_des_chaines_de_reference_DEMOCRAT (visited on 10/13/2024).

— (2021). "Le corpus Democrat et son exploitation. Présentation". In: *Langages* 4, 11–24.

Langlais, Pierre-Carl (2021). *Fictions littéraires de Gallica / Literary Fictions of Gallica*. Version 1. Zenodo. 10.5281/zenodo.4751204.

Le, Quoc V. and Tomás Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*, 1188–1196. 10.48550/arXiv.1405.4053.

Leblond, Aude (2022). *Corpus Chapitres*. Version 1.0.0. 10.5281/zenodo.7446728.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 188–197. 10.18653/v1/D17-1018.

Lotman, Yuri (1977). *The Structure of the Artistic Text*. Michigan Slavic Contributions 7. Michigan University Press.

Luo, Xiaoqiang and Sameer Pradhan (2016). "Evaluation Metrics". In: *Anaphora Resolution: Algorithms, Resources, and Applications*. Ed. by Massimo Poesio, Roland Stuckardt, and Yannick Versley. Springer, 141–163. 10.1007/978-3-662-47909-4_5.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 55–60. 10.3115/v1/P14-5010.

Manovich, Lev (2016). "The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics". In: *Journal of Cultural Analytics* 1 (1). 10.22148/16.004.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot (2020). "CamemBERT: A Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 10.18653/v1/2020.acl-main.645.

Maynard, Diana, Kalina Bontcheva, and Isabelle Augenstein (2017). "Named Entity Recognition and Classification". In: *Natural Language Processing for the Semantic Web*. Springer International Publishing, 25–36. 10.1007/978-3-031-79474-2_3.

Milli, Smitha and David Bamman (2016). "Beyond Canonical Texts: A Computational Analysis of Fanfiction". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2048–2053. 10.18653/v1/D16-1218.

Moretti, Franco (2000). "Conjectures on World Literature". In: *New Left Review*.

Naguib, Marco, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, and Olga Seminck (2022). "Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS (Male and Female Novelists: An Automatic Study of Gender of Authors and Their Characters)". In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, 66–77. https://aclanthology.org/2022.jeptalnrecital-humanum.8 (visited on 10/13/2024).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent

Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (visited on 10/13/2024).

Perri, Vincenzo, Lisi Qarkaxhija, Albin Zehe, Andreas Hotho, and Ingo Scholtes (2022). "One Graph to Rule Them All: Using NLP and Graph Neural Networks to Analyse Tolkien's Legendarium". In: *arXiv preprint*. 10.48550/arXiv.2210.07871.

Piper, Andrew and Sunyam Bagga (2022). "A Quantitative Study of Fictional Things". In: *Proceedings of Computational Humanities Research*, 268–279. https://ceur-ws.org/Vol-3290/long_paper1576.pdf (visited on 10/13/2024).

Piper, Andrew, Richard Jean So, and David Bamman (2021). "Narrative Theory for Computational Narrative Understanding". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 298–311. 10.18653/v1/2021.emnlp-main.26.

Poesio, Massimo, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal (2023). "Computational Models of Anaphora". In: *Annual Review of Linguistics* 9, 561–587. 10.1146/annurev-linguistics-031120-111653.

Rockwell, Geoffrey and Stéfan Sinclair (2016). *Hermeneutica: Computer-assisted Interpretation in the Humanities*. MIT Press.

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

Schaeffer, Jean-Marie (1989). *Qu'est-ce qu'un genre littéraire?* Poétique. Seuil.

Schlegel, Friedrich, August Wilhelm Schlegel, August Ferdinand Bernhardi, and Wilhelm Dilthey (1996). *Critique et herméneutique dans le premier romantisme allemand : Textes de F. Schlegel, F. Schleiermacher, F. Ast, A.W. Schlegel, A.F. Bernhardi, W. Dilthey*. Trans. by Denis Thouard. Opuscules. Presses universitaires du Septentrion. https://books.openedition.org/septentrion/95397 (visited on 01/23/2024).

Schmid, Wolf (2010a). *Mental Events. Changes of Mind in European Narratives from the Middle Ages to Postrealism*. Hamburg University Press. 10.15460/HUP.215.

— (2010b). *Narratology. An Introduction*. De Gruyter.

Schöch, Christof (2017). "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". In: *Digital Humanities Quarterly* 11 (2). https://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (visited on 10/23/2024).

Seminck, Olga, Philippe Gambette, Dominique Legallois, and Thierry Poibeau (2022). "The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature". In: *Journal of Cultural Analytics* 7 (3). 10.22148/001c.37588.

Silge, Julia and David Robinson (2017). *Text Mining with R: A Tidy Approach*. http://repo.darmajaya.ac.id/5417/ (visited on 10/13/2024).

Sims, Matthew, Jong Ho Park, and David Bamman (2019). "Literary Event Detection". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3623–3634. 10.18653/v1/P19-1353.

Sobchuk, Oleg and Artjoms Šeļa (2024). "Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction". In: *Humanities and Social Sciences Communications* 11. 10.1057/s41599-024-02933-6.

Soni, Sandeep, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens, and David Bamman (2023). "Grounding Characters and Places in Narrative Texts". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 11723–11736. 10.18653/v1/2023.acl-long.655.

Sprugnoli, Rachele and Sara Tonelli (2016). "Novel Event Detection and Classification for Historical Texts". In: *Computational Linguistics* 45 (2), 229–265. 10.1162/coli_a_00347.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (2012). "brat: A Web-based Tool for NLP-assisted Text Annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107. https://aclanthology.org/E12-2021 (visited on 10/13/2024).

Toro Isaza, Paulina, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang (2023). "Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children's Fairy Tales". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6509–6531. 10.18653/v1/2023.acl-long.359.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). "Llama 2: Open Foundation and Fine-tuned Chat Models". In: *arXiv preprint*. 10.48550/arXiv.2307.09288.

Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 1–33.

Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The Transformation of Gender in English-language Fiction". In: *Journal of Cultural Analytics* 3 (2). 10.22148/16.019.

van Cranenburgh, Andreas and Frank van den Berg (2023). "Direct Speech Quote Attribution for Dutch Literature". In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 45–62. 10.18653/v1/2023.latechclfl-1.6.

van Zundert, Joris, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Mussmann (2022). "What Do We Talk About When We Talk About Topic?" In: *Proceedings of Computational Humanities Research*. https://ceur-ws.org/Vol-3290/short_paper5533.pdf.

van Zundert, Joris, Andreas van Cranenburgh, and Roel Smeets (2023). "Putting Dutchcoref to the Test: Character Detection and Gender Dynamics in Contemporary Dutch Novels". In: *Proceedings of Computational Humanities Research*, 757–771. https://ceur-ws.org/Vol-3558/paper9264.pdf (visited on 10/13/2024).

Vanni, Laurent, Melanie Ducoffe, Carlos Aguilar, Frederic Precioso, and Damon Mayaffre (2018). "Textual Deconvolution Saliency (TDS) : A Deep Tool Box for Linguistic Analysis". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 548–557. 10.18653/v1/P18-1051.

Vianne, Laurine, Yoann Dupont, and Jean Barré (2023). "Gender Bias in French Literature". In: *Proceedings of Computational Humanities Research*, 247–262. https://ceur-ws.org/Vol-3558/paper2449.pdf (visited on 10/13/2024).

Vishnubhotla, Krishnapriya, Frank Rudzicz, Graeme Hirst, and Adam Hammond (2023). "Improving Automatic Quotation Attribution in Literary Novels". In: *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics*, 737–746. 10.18653/v1/2023.acl-short.64.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston (2013). *Ontonotes Release 5.0 LDC2013T19*. 10.35111/xmhb-2b84.

Woloch, Alex (2003). *The One vs. the Many*. Princeton University Press. http://www.jstor.org/stable/j.ctt7srp4 (visited on 10/13/2024).

Yu, Bei (2008). "An Evaluation of Text Classification Methods for Literary Study". In: *Literary and Linguistic Computing* 23 (3), 327–343. 10.1093/llc/fqn015.

Zehe, Albin, Leonard Konle, Lea K. Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer (2021). "Detecting Scenes in Fiction: A new Segmentation Task". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3167–3177. 10.18653/v1/2021.eacl-main.276.

Zhang, Weiwei, Jackie Chi Kit Cheung, and Joel Oren (2019). "Generating Character Descriptions for Automatic Summarization of Fiction". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (1), 7476–7483. 10.1609/aaai.v33i01.33017476.

A. Confusion Matrices for BookNLP-fr-based Models

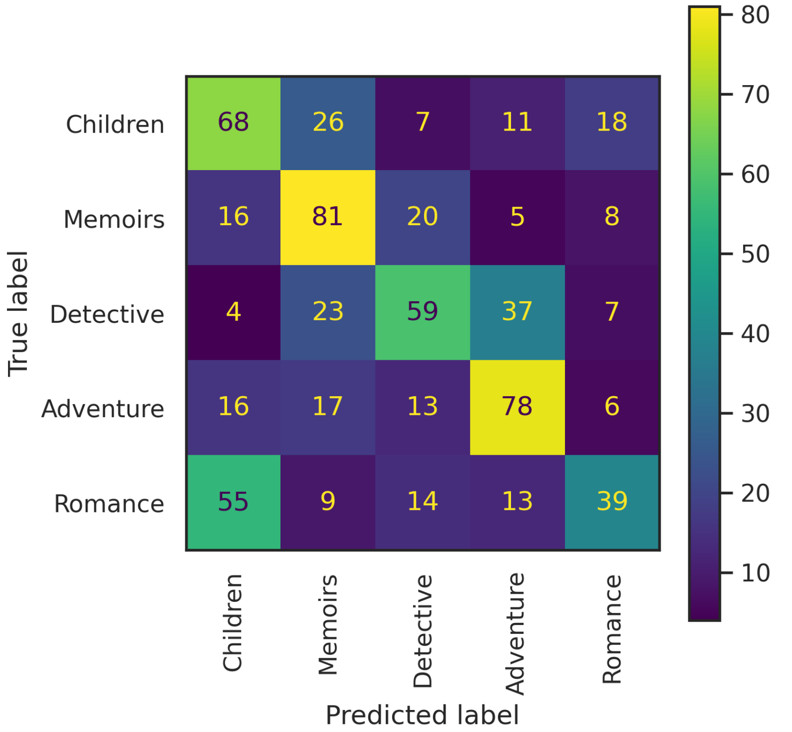

Figure 7: Confusion matrix for ADJ features.

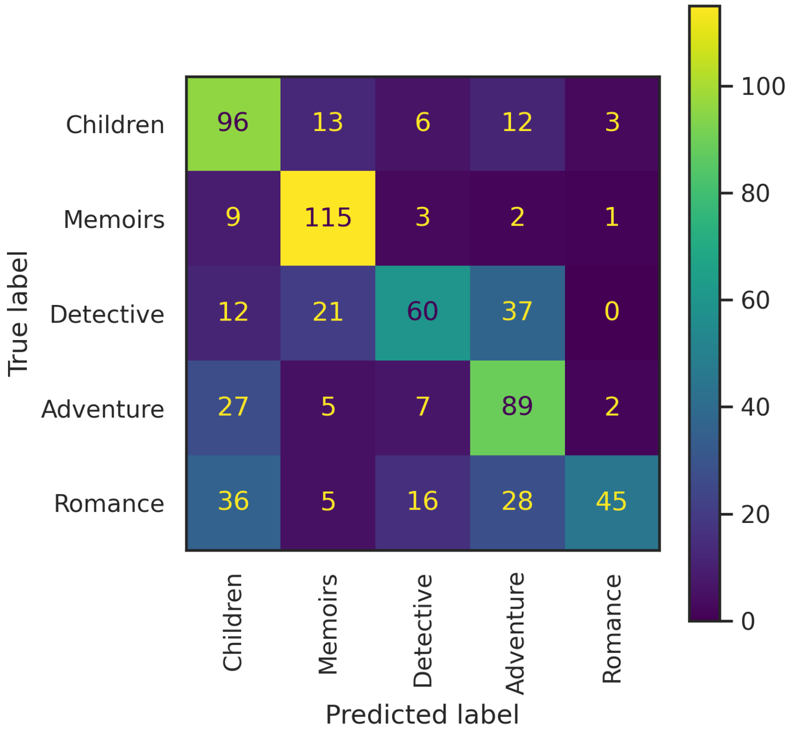

Figure 8: Confusion matrix for AGENT features.

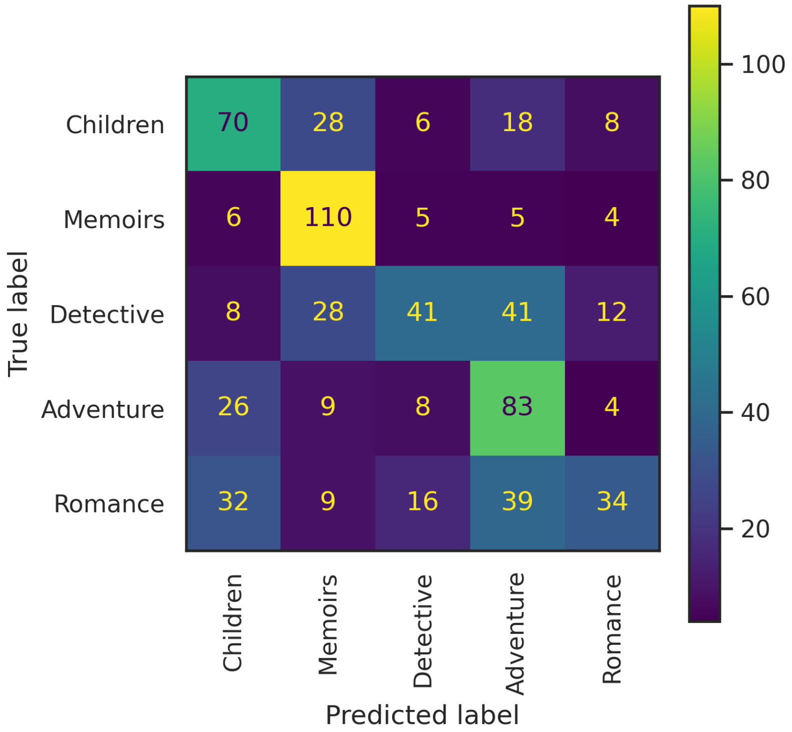

Figure 9: Confusion matrix for PATIENT features.

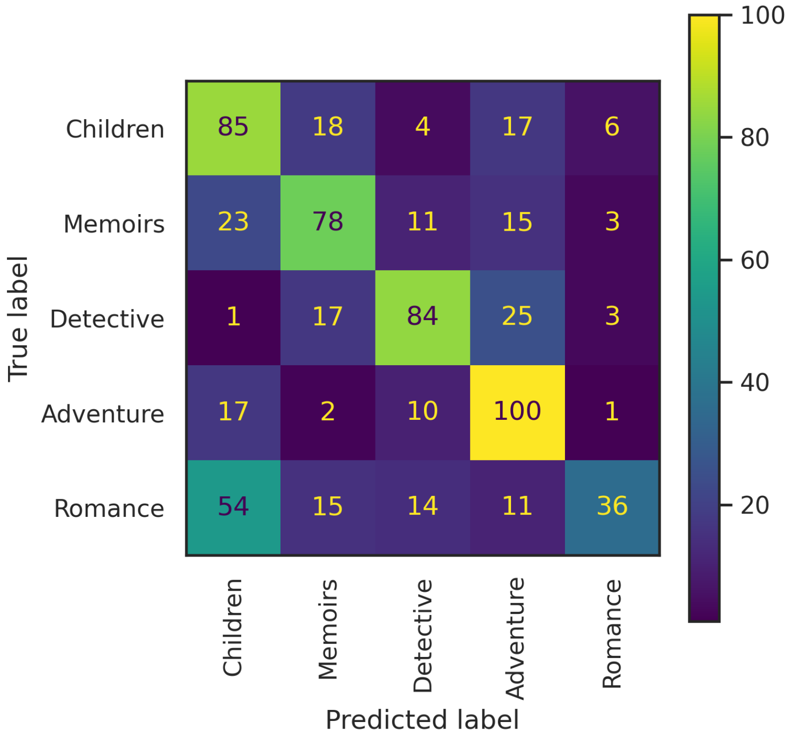

Figure 10: Confusion matrix for FAC features.

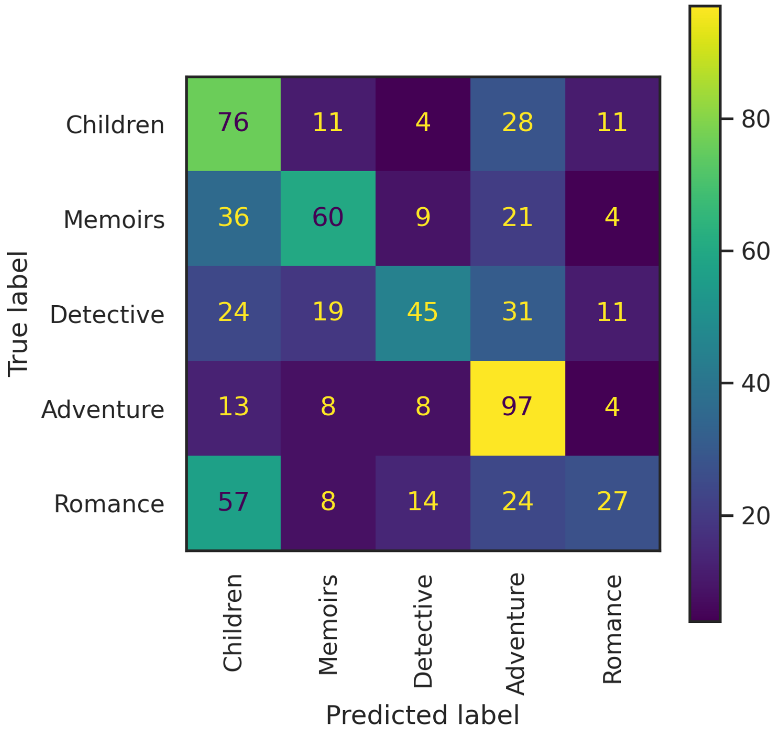

Figure 11: Confusion matrix for GPE features.

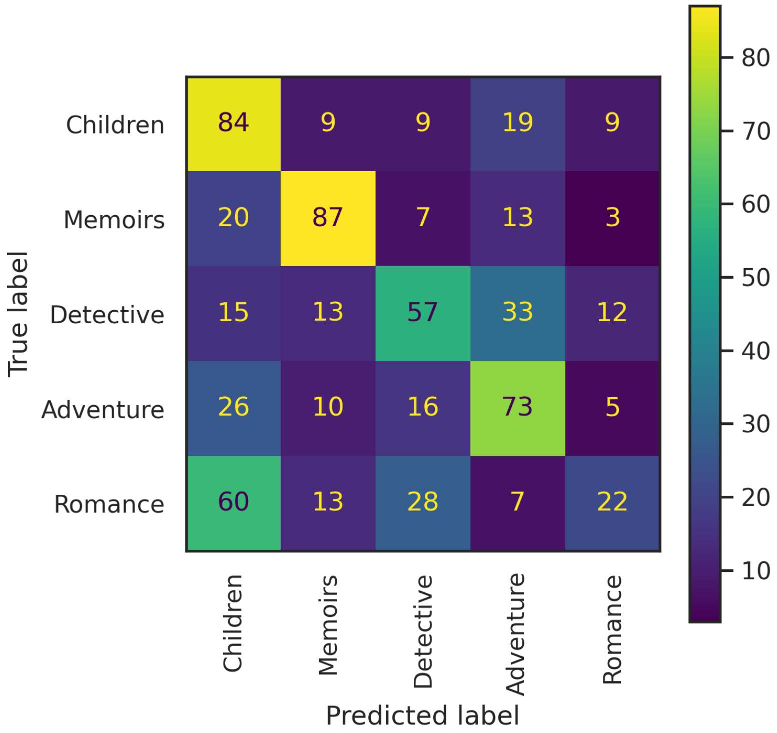

Figure 12: Confusion matrix for TIME features.

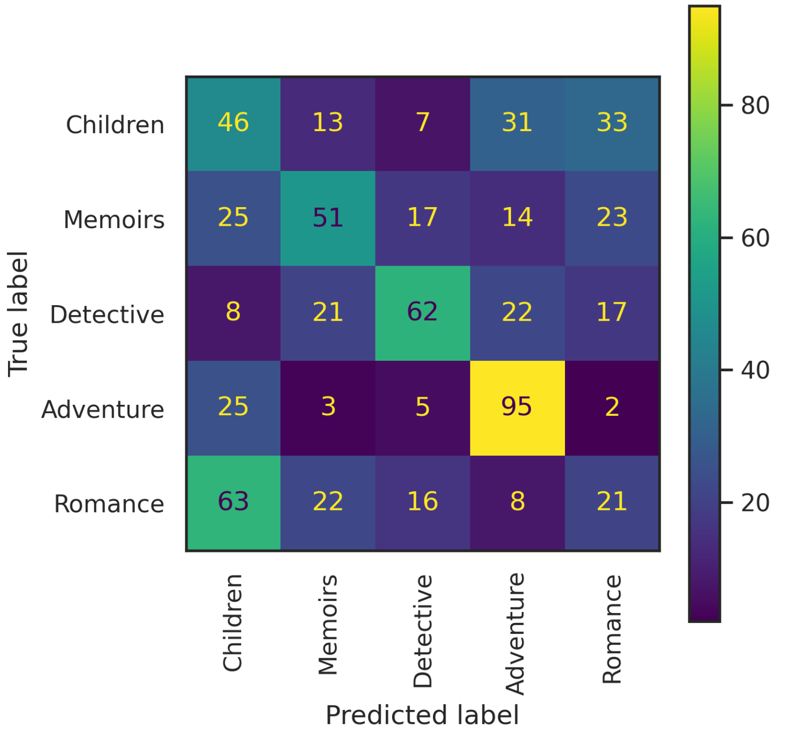

Figure 13: Confusion matrix for VEH features.

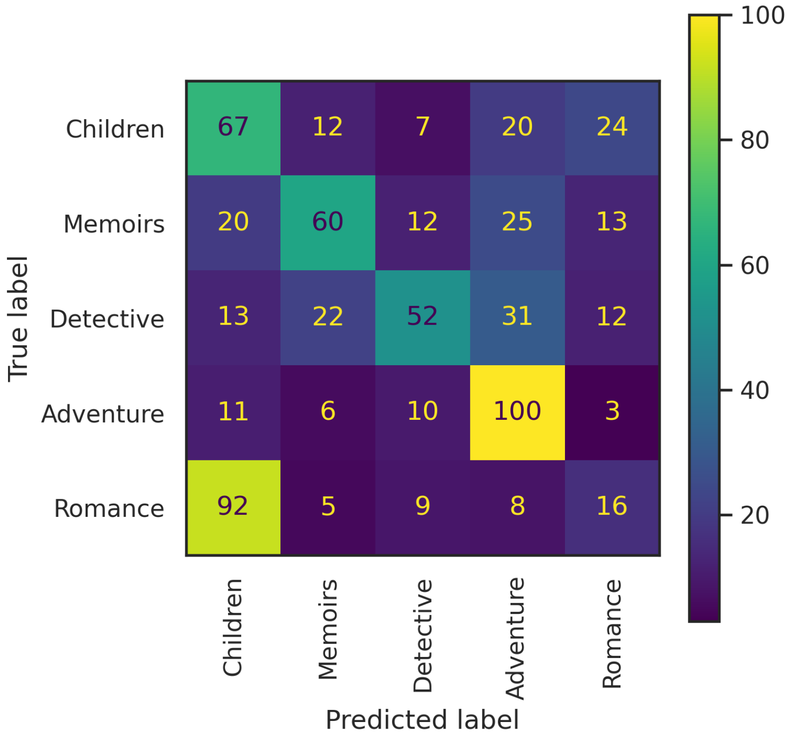

Figure 14: Confusion matrix for LOC features.