CLAUSE-ATLAS: A Corpus of Narrative Information to Scale Up Computational Literary Analysis

Enrica Troiano, Piek Vossen
Vrije Universiteit Amsterdam, Computational Linguistics and Text Mining Lab
{e.troiano, p.t.j.m.vossen}@vu.nl

Abstract

We introduce CLAUSE-ATLAS, a resource of XIX and XX century English novels annotated automatically. This corpus, which contains 41,715 labeled clauses, allows to study stories as sequences of eventive, subjective and contextual information. We use it to investigate if recent large language models, in particular `gpt-3.5-turbo` with 16k tokens of context, constitute promising tools to annotate large amounts of data for literary studies (we show that this is the case). Moreover, by analyzing the annotations so collected, we find that our clause-based approach to literature captures structural patterns within books, as well as qualitative differences between them.

Keywords: narrative theory, LLM-based annotation, literary resources, subjectivity, events, ChatGPT

1. Introduction

Written stories run through the course of cultural evolution (Boyd, 2018; Sugiyama, 2005). Some reproduce societal norms and values (Wiessner, 2014; Tirrell, 1990), while others create worlds that have no correspondence to reality. In either case, stories transfer a great deal of knowledge from writers to readers, often involving the communication of intricate plots, the use of creative linguistic devices, and the engagement of multiple mental representations – for example about the psychological depth of characters (Currie, 2009; Boyd, 2017), the network of their actions (Adams, 1989), and the geography and historical context in which they live (Habermann and Kuhn, 2011; De Groot, 2009). Thanks to these aspects, narratives represent interesting data for natural language processing (NLP), but they also raise an important question: what does it mean to understand a story?

Typically, studies model this human skill with language technologies focused on the basic elements of stories, like events (Sims et al., 2019; Toro Isaza et al., 2023), characters (Bamman et al., 2020; Konovalova and Toral, 2022), or their features – e.g., gender (Oka and Ando, 2020), emotions (Mohammad, 2011; Kim and Klinger, 2018), personality (Pizzolli and Strapparava, 2019). Following up on such works, we describe a new approach to analyze novels, grounded in a theory of narrative comprehension.

We introduce the corpus CLAUSE-ATLAS[1], comprising six novels in English. We annotate the novels with three types of narrative information documented in the theoretical framework of Berman (1997). Elaborating on Labov's idea that narrative construction revolves around tensed clauses (1972), Berman proposes a taxonomy composed by narrative elements (clauses that report eventive aspects in a story), evaluative elements (those that express personal perspectives, for example the motivations and mental states of the characters), and informative elements (any additional background information). We label the clauses in our corpus accordingly (an illustration of the task is in Figure 1), producing a "clause atlas" for literature analysis.

When applied to entire books, such an approach allows to describe literature in innovative ways, specifically in terms of how stories unfold. This could hardly be done with existing resources, which provide annotations for *sparse* text (e.g., only the portions that denote characters) of relatively short *samples* of books (Bamman et al., 2014; Vala et al., 2016). Instead, we obtain annotations for entire literary works (i.e., no single token in them is left unlabeled), using a pipeline based on Large Language Models (LLMs). More precisely, we employ classifiers implemented through ChatGPT prompting strategies. The achievements of ChatGPT in many tasks, like stance, topic, and frame detection (Gilardi et al., 2023), raise expectations about its potential use for computational literary research. This tool could give the unprecedented possibility of studying texts in sizes beyond what can be easily analyzed by humans. Its narrative understanding abilities, however, still require careful assessment.

Therefore, we use CLAUSE-ATLAS to answer the following research questions: (RQ1) Is ChatGPT a good annotator of narrative constructs? (RQ2) Can its annotation of narrative, evaluative, and informative clauses reveal structural qualities of novels? We find that ChatGPT is indeed a reliable literature annotator. By testing it in a zero- and few-shot classification scenarios, and by comparing its output to that of humans, we show that differ-

[1] CLAUSE-level Annotation of Texts for Literary AnalysiS is available at `https://huggingface.co/datasets/troianea/CLAUSE-ATLAS`.

LREC-COLING 2024, pages 3283–3296
20-25 May, 2024.

Achilles withdrew from the battle,
filled with anger and resentment
for the conflict with Agamemnon.

Figure 1: Our clause-level annotation task, with narrative, evaluative, and informative clauses.

ent prompts cause heterogeneity in the annotation outputs, comparable to the different annotation behaviors found among people. Further, we provide qualitative evidence that our theoretically-motivated annotation schema allows to perform in-depth analyses of novels.

In sum: (1) we clarify the feasibility of a methodology based on ChatGPT for literary data creation, which scales beyond the limits of current human-bound book annotations; (2) we release a corpus annotated automatically and (partly) by humans, which resonates with Berman's clause-based taxonomy at a higher and more conceptual level of understanding than the elemental building blocks of stories (characters and/or sequences of events); (3) we deliver an analysis of novels that captures their structural qualities, both at a meta (book) level, and at finer-grained scales.

2. Related Work

Many collections of literary texts have been annotated by humans to reason about narrativity. Given the successful application of Large Language Models for text analysis, this manual labor could soon be taken over by automatic systems like ChatGPT.[2] Below, we review available resources for literary analysis and recent research based on ChatGPT.

Annotating Narratives. Studying narratives poses a crucial data-related problem: how to find substantial amounts of labeled texts. The majority of resources comprise constrained forms of short plots, called scenarios (Regneri et al., 2010; Modi et al., 2016), but annotations of entire books are rarer, because they require a great deal of human effort. Indeed, the size of resources of this type is substantial when the annotations qualify books at a meta level, and it drops exponentially as the granularity of the annotation task increases. For instance, the CONLIT dataset spans 2,700 works of contemporary literature associated to labels on genre, category, author, and publication date (Piper, 2022), while Bamman et al.'s corpus (2014), having coreference data, covers five novels, later extended to 100, of which only samples are annotated (Bamman et al., 2020).

Several corpora contain information about characters, particularly concerning their features or interactions. For example, LiSCU accompanies literary pieces with descriptions of the characters that appear in them (Brahman et al., 2021), the dataset curated by Flekova and Gurevych (2015) documents the personality traits of 298 book characters, and others mark speakers (Stymne and Östman, 2020; Ek et al., 2018) and their speech (Papay and Padó, 2020; Vishnubhotla et al., 2022). Datasets have also been created to study narrative events (Sims et al., 2019; Meehan and Piper, 2022) and sequences thereof (Reiter, 2015). Among these, the one resource with which ours has some points of contact is the German corpus presented in Vauth et al. (2021): it contains verbal phrases annotated with fine-grained event categories, including facts in the narrated world and internal to the characters.

ChatGPT for Text Annotation. ChatGPT is a LLM-based assistant. Introduced in November 2022, it was trained in a framework of Reinforcement Learning from Human Feedback where it acquired its language proficiency from massive volumes of human annotations (Schulman et al., 2022). Among its many capabilities is text classification in zero-shot scenarios (i.e., requiring no additional training), where ChatGPT has competitive results on tasks like genre identification (Kuzman et al., 2023), fact-checking (Hoes et al., 2023), and detection of pragmatic components of apologies (Yu et al., 2023a).

Users interact with ChatGPT via their prompts, i.e., task instructions that enforce specific rules about the quality and quantity of the generated outputs. Prompts are an efficient way of automatizing processes because they are written in natural language, but minor variations in their wording can elicit extremely different outputs. For this reason, the careful engineering of prompts has become an important requirement for the application of models like ChatGPT in research, together with the need for rigorous comparisons between automatic and human-generated annotations (Reiss, 2023).

Key Differences. Unlike the reviewed research, we concentrate on an abstract narrative layer, concerned with the distinction between eventive information (i.e., happenings in the external world) and subjective information (i.e., the private sphere of characters). In this sense, our approach is close to analyses of story structures, like the annotation task of Ouyang and McKeown (2014) assigning textual clauses to a function of narrative progression (e.g., complicating action, resolution, coda). The phenomena we investigate, however, are more similar to those of Wiebe (1994), who looked into the subjective sentences of novels to find regularities in how authors manipulate the psychological point of view of characters (e.g., thoughts, perceptions, inner states). We venture into some of the direc-

[2] https://chat.openai.com

Event: All that happens outside the characters. Clauses of this sort depict objective facts and the progression of the story plot, like activities, achievements, actions.

Subj.Ex.: What happens in the characters' "mind", like thoughts, perceptions, emotions, memories, personal interpretations of the outside world, beliefs, motives.

Cont.Inf.: Additional information that helps the reader understand the story. These clauses describe characters, narrative world, historical context, cultural background, or relationships between events.

Table 1: Label definition in our schema. Subj.Ex.: subjective experience. Cont.Inf.: contextual information.

tions left open by this work that has not released any resource.

For that, we take on the idea that ChatGPT excels at subjective tasks such as hate speech detection, for which its understanding turns out on par with humans (Huang et al., 2023), and stance detection, where it surpasses other state-of-the-art classification models (Zhang et al., 2022). Hence, we bring it in the NLP subfield of computational literary analysis, still lacking resources thoroughly annotated with the assistance of LLMs.

3. Layers of Narrative Understanding

Our goal is to test if ChatGPT is a good literature annotator, and whether its output, as based on a theoretical schema, grasps literary qualities of novels. We pursue these objectives starting from two observations. First: from a literary text, readers learn or infer a profusion of narrative information (Iser, 1979), e.g., an event happens, it bears a certain significance for a character, it determines future actions or beliefs. Second: this information emerges at various linguistic granularities (Kintsch, 1998). As opposed to shorter productions (e.g., online posts), books tend to be more structured: sentences form the paragraphs of a chapter, chapters compose volumes, and so on – a story's understanding happens at all such granularities.

These aspects represent two operational challenges for computational narrative analysis, in terms of what information to investigate, and at what linguistic level that should be done.

Annotation Schema. The information that we consider relates to the readers' ability to distinguish between things that occur within a character (e.g., stances, motivations, perceptions) and objective facts in the narrative world (e.g., achievements, actions, interactions). These dimensions are well formalized by Berman (1997), according to whom narratives are made of three types of expressions, which carry either a narrative, an evaluative, or an informative function.

Narrative elements report what happens in the story, and are related to one another by ties of temporal sequentiality; evaluative elements reveal the characters' or narrator's personal perspectives regarding motivations, emotions, mental states, interpretations, judgements; informative texts give background information about the characters, the circumstances in which events take place, or any other detail that the readers might need or want to know. In our study, these three constructs map to the annotation labels "event", "subjective experience", and "contextual information", defined as in Table 1. Note that we refrain from treating events in any formal sense. We consider them as encompassing all kinds of factual happenings.

Annotation Unit. A way to apply our labels to novels would be to annotate tokens or phrases. For example, one could identify emotion words as indicators of subjective experiences (e.g., as in Kim and Klinger, 2018). However, the information carried by tokens is crucially context-dependent – e.g., "*see*" might indicate a perceptive process ("*I see you*") or be used metaphorically ("*I see what you mean*"). Moreover, given our goal to cover entire books, using tokens as annotation units is expensive: it requires to devise a compositional function to map individual annotations into a final label for larger spans of text, and to make a comprehensive list of the phenomena signaling a subjective experience, an event, and contextual information.

We decide instead to focus on grammatical word sequences, i.e., clauses (one clause → one label). Also the information present in a clause can depend on the surrounding context. For example, the text "*that is the tallest duck in the world*" describes an entity, thus carrying contextual information, but it might be a subjective experience as well, if preceded by "*Felix thought:*". Still, despite this challenge, clauses are a good compromise between the granularity of annotation of tokens, sentences, or paragraphs. They provide propositional context to resolve token ambiguities, and they likely realize just one of the three information units (Berman, 1997), while larger resolutions might express many (see the sentence in Figure 1). Lastly, should we find that our annotation captures some qualitative patterns of narrativity, clauses could be leveraged for the computational processing of books, since they can be encoded by current technologies more easily than longer texts (e.g., chapters).

4. Data and Preprocessing

Because our study is the first to venture a full-scale, fully-automatic annotation for narrative analysis, we gain insight into the promise of this approach using a handful of books.[3] We choose six from the XIX and XX centuries: three children's fictions (Alice's Adventures in Wonderland, Peter Pan, The Adventures of Pinocchio) and three cult novels (Frankenstein, Pride and Prejudice, The Great Gatsby).

To identify clauses, we divide each book into sentences while keeping track of the paragraph and chapters they belong to.[4] Sentences do not naturally contain markers of clause boundaries, and to the best of our knowledge there are no ready-to-use tools for clause identification. Hence, we achieve this goal through ChatGPT, accessing the LLM (`gpt-3.5-turbo` with 16k tokens of context) via the official OpenAI's API. We set the model's temperature to 0, and leverage its *function calling*[5] capability to structure the outputs as JSON objects. We identify all grammatical clauses in a sentence using a prompt that proved to elicit satisfactory results.[6] Such a prompt concatenates our description of the clause segmentation task with the sentence on which it should be accomplished. Splitting a book into clauses thus requires to make as many API calls as the number of sentences in each book.

We obtain this way the data ready for annotation.

5. Annotation

The creation of CLAUSE-ATLAS took place between August and October 2023. It comprised two tasks: 1. mapping a clause to a label (event, subjective experience, contextual information); 2. recognizing experiencers for all (and only) the subjective experience clauses. In what follows, we detail how we annotated the texts, and we describe the final corpus.

5.1. Setup

As shown in Figure 2, we carry out our study in an automatic and a human-based setups. In both, the assignment is to read a paragraph and analyze the given clauses in the order in which they appear.

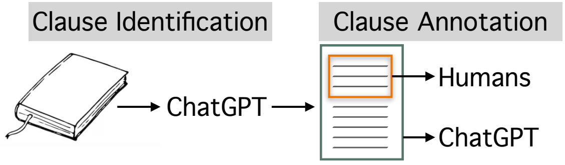

Figure 2: Our annotation setup. The clauses of a book found by ChatGPT are annotated automatically and by humans. The latter examine the clauses in the first chapter only.

Automatic. We rely once more on ChatGPT[7], and we perform the tasks sequentially, such that the output of task 1. serve as an input for task 2. We collect ChatGPT's responses by feeding it with a prompt that concatenates our task instructions and the textual input to analyze, i.e., all clauses of a paragraph for 1., and all previously-labeled clauses of a paragraph for 2. Collecting responses in each task involves n API calls, where n = the number of paragraphs in a book. We repeat this process for all books.

To address the labeling task (1.), we experiment with prompts characterized by three different instructions, at increasing degrees of detail. The first prompt (P1) corresponds to a zero-shot classifier. It provides definitions of subjective experiences, events, and contextual information. The second (P2) provides definitions and examples – hence corresponding to the few-shot scenario. The third (P3) reports examples (i.e., it is also a few-shot classifier), more detailed definitions, and an illustration of the expected output. We use P1, P2, and P3 to gather three annotations for the books.

Concerning the experiencer recognition task (2.), our instructions include a list of characters that we manually extracted from the Wikipedia page of each book, as well as the option "Other". We ask ChatGPT to choose (subjectivity) experiencers from there, to increase the chance of obtaining character names instead of pronouns. We perform this task only with the prompt that proves the most reliable (see Section 6.1).

Human-based. The human annotation is conducted by three Master's students proficient English (one is an English native speaker), and with some background in NLP, literary studies and linguistics.

We converge on the understanding of the labels during a session of training, and ask them to annotate only the first chapter of Alice's Adventures in Wonderland, The Adventures of Pinocchio, and The Great Gatsby, to make their workload manageable. After the training session, the annotators work independently to complete their job in (approx.) 24 hours, spread over a period of 2 weeks.

We render the labeling task coherent with the automatic variant, through guidelines that correspond to the most detailed prompt (P3). As for task 2.,

[3] Extracted from https://www.gutenberg.org.

[4] We treat as a paragraph any block of text delimited by empty lines.

[5] https://openai.com/blog/function-calling-and-other-api-updates

[6] Prompts and annotation guidelines can be found at https://huggingface.co/datasets/troianea/CLAUSE-ATLAS. In Appendix A, we report a manual analysis of the quality of clauses.

[7] Model parameters as in Section 4.

	Alice	Peter	Pinoc.	Frank.	Pride	Gatsby
Ch.	12	17	36	28*	61	9
Pa.	787	1,653	1,716	764	2,060	1,610
Sn.	1,508	3,168	3,263	3,347	5,957	3,441
Cl.	3,372	6,287	5,584	7,928	12,419	6,125
Tk.	34,091	59,868	50,364	90,965	150,943	61,502

Table 2: Statistics on CLAUSE-ATLAS. Ch.: chapters. Pa.: paragraphs. Sn.: sentences. Cl: clauses. Tk.: tokens (tokenization performed with tiktoken[10]). *: 24 chapters + 4 letters.

humans choose characters involved in a subjective experience as soon as a clause is labeled as such. Characters are picked from the list we provide.

5.2. CLAUSE-ATLAS

The corpus so constructed encompasses 41,715 clauses (distributed as in Table 2), which are associated with 6 layers of labels, three from humans (covering the first chapter of three books), and three from ChatGPT. A comparison of the distribution of this annotation is illustrated in Figure 3.

Focusing on the annotations of the prompt that we use to carry out the analysis of stories (Section 7), we find that books contain 146 experiencers on average (7 times more than the average list of characters we provided for the corresponding annotation).[8] The narrator always appears among the three most common experiencers besides the major protagonists, both when the narration is in first person (e.g., in Frankenstein, where the narrator is an experiencer 1,939 times, and The Monster 687 times) and when the text is written in third person (e.g., in Peter Pan, Peter is an experiencer 561 times, Wendy Darling 459 times, and the narrator 403 times).

6. Is ChatGPT an Appropriate Literature Annotator?

We now assess the degree to which ChatGPT has an internal model of the theory we follow. To answer our first research question, we consider inter-annotator agreement within and between our two annotation setups. First, we examine whether ChatGPT's understanding of subjective experiences, events, and contextual information is stable across prompts (Section 6.1). Next, we compare its outputs to those of humans (Section 6.2).

6.1. Output Stability

If ChatGPT has a representation of the concepts encoded in our schema, prompts that describe those concepts at different levels of abstraction should yield consistent responses (and high agreement, accordingly). Contrary to this idea, previous work has found that minimal prompt variations can elicit incompatible outputs (Kuzman et al., 2023). We clarify the issue with respect to our literary task: we consider each prompt as a different annotator, and we analyze their labeling choices for the 41,715 clauses in CLAUSE-ATLAS using Fleiss' κ (1971).

Agreement turns out moderate: κ is 0.57, with the highest score being achieved on The Adventures of Pinocchio (κ = 0.60), and the lowest on Pride and Prejudice (κ = 0.51). We further observe that agreement does not reflect the complexity of the language in a book – e.g., it is comparable between a kids' book such as Alice's Adventures in Wonderland and The Great Gatsby (κ = 0.54 and 0.58, respectively).

These results give a first evaluation of ChatGPT's output consistency, but they likely underestimate agreement in our task. The three outputs are skewed towards the choice "subjective experience" (as illustrated for P1 in Figure 4, reporting the distribution of labels in each novel). This causes a higher expected agreement, and penalizes the overall κ score. As a solution, we analyze raw agreement counts. In comparison to the κ ratios, these scores suggest that ChatGPT's outputs are substantially more stable. For more than half the clauses, P1, P2 and P3 are in complete agreement: 11,350 of such clauses receive three "subjective experience" judgments, followed by the clauses marked as "contextual information" (7,155), and a smaller number of instances where the prompts spurred perfect agreement on "event" (6,144). For 15,930 clauses there is a majority vote, and only on 1,136 clauses the automatic annotators pick three different labels.

In sum, ChatGPT's answers obtained with different instructions converge more often than not, but they are not always identical. We therefore need to establish if this variability is acceptable, by observing it against humans'.

6.2. Comparison to Humans

In CLAUSE-ATLAS, 992 clauses are annotated by both humans and ChatGPT (i.e., the first chapter of Alice's Adventures on Wonderland, The Adventures of Pinocchio and The Great Gatsby). To investigate the annotation quality on these items, we adopt traditional inter-annotator agreement measures (which summarize agreement as a single aggregated score), as well as the CrowdTruth frame-

[8] CLAUSE-ATLAS experiencers exceed the main characters in the list given during annotation due to the frequent selection of the option "Other".

[9] https://github.com/openai/tiktoken

[10] https://github.com/openai/tiktoken

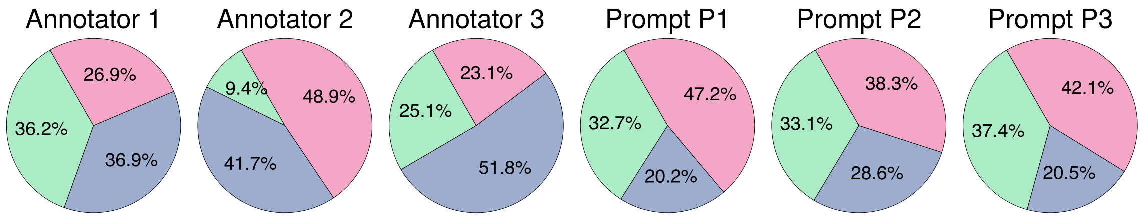

Figure 3: Annotation comparison: events, subjective experiences, and contextual information.

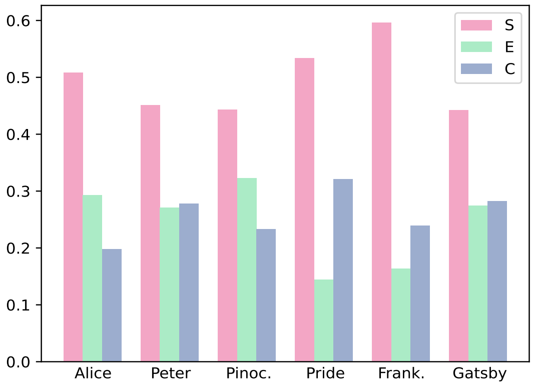

Figure 4: Relative frequency (y-axis) of labels per book, from the annotation of P1. S: subjective experience. E: event. C: contextual information.

work[11] (Dumitrache, 2019), which breaks down agreement with an ensemble of metrics for workers, annotation units and labels, separately. These components are represented as weighted vectors (e.g., annotation units that caused substantial disagreement are given lower weights, like workers who frequently disagree with their peers), and the corresponding metrics (ranging from 0 to 1) are estimated through cosine similarities.

Notably, the unit-quality-score reveals if some annotation units are ambiguous, and the worker-quality-score helps identify untrustworthy workers. The possibility of teasing apart the two aspects is useful for subjective tasks like ours, where disagreement might not be due to the annotators' unreliability but to their different readings of clauses that fit many (plausible) interpretations – as a matter of facts, some aspects of story comprehension, like the attitudes of characters, might not be openly named in a clause. Hence, we base this analysis on the following criterion: an ideal scenario is not necessarily one where ChatGPT reaches high inter-annotator agreement with humans (since we can expect humans themselves to make different annotation choices), but it is one in which such agreement approximates that among humans.

Quantitative Analysis. Figure 3 shows the diversity of annotations between the two setups. We see, e.g., that the prompt coherent with the human guidelines (P3) produced outputs more similar to the other prompts than to the students. Moreover, humans preferred contextual information, contrary to their automatic counterparts. Correspondingly, we find differences in the Fleiss' scores: P1, P2 and P3 obtain a κ of 0.58, while κ = 0.32 for the humans alone.

Thus, as anticipated, humans achieve poor agreement, but their κ compares to the agreement computed on the six annotators all together (0.33), and to the average κ (0.31) found on different combinations of three annotators extracted from the two setups. These findings confirm our desideratum about the ChatGPT-humans similarity.

We further corroborate them via agreement counts. On average, pairs of human and automatic annotators agree on 49% of the data, while human pairs agree more often (on 56% of the items). The distributions of the pairwise percentage agreements in the two cases, however, is not significantly different (p-value>0.05 with a Mann-Whitney U test (Mann and Whitney, 1947)).

Interestingly, these proportions increase by computing inter-annotator agreement on the 393 items where humans had 3 unanimous judgments. Pairs of human-automatic annotators agree on 62% of such items, 13 points more than the raw counts.[12] This also happens by using perfect agreement in one annotation setup as a condition to analyze agreement in the other: pairs of ChatGPT-based annotators provided identical judgments on 73% of the data, and on the items where the three humans chose the same label (393 clauses), they agree 76% of the times. Vice versa, human pairs achieve agreement on 58% of the clauses on which P1, P2 and P3 picked the same label.

We further assess the annotation quality in CrowdTruth. Here, humans and ChatGPT are clearly separated, as the former obtain workers-quality-scores between 0.41 and 0.50, while Chat-

[11] Following the code at https://github.com/CrowdTruth/CrowdTruth-core/.

[12] The distribution of pairwise agreement on all 992 clauses differs significantly from that conditioned on the subsample where humans reach perfect agreement (p-value = 0.0003 with Mann-Whitney U test).

GPT ranges from 0.57 (for P3) to 0.59 (for P1).

Put together, these results provide three pieces of evidence. First, agreement between the automatic and the human annotators is low but it is on par with the humans' alone. Second, ChatGPT is not less reliable than humans. The different instruction wordings resulted in label variability, but that variability is found among people to an even greater extent; further, on the subset of clauses that humans agreed upon, also the agreement among ChatGPT outputs is substantial. Third, the output quality of one prompt (P1) surpasses all others, receiving the best worker-quality-score.[13] We focus on the annotations generated with this zero-shot classifier to investigate the structure of novels in CLAUSE-ATLAS (Section 7).

Qualitative Analysis of Disagreements. To grasp what underlies disagreements in our task, we manually investigate 540 clauses with a unit-quality-score smaller than 0.5. In the CrowdTruth framework, such a metric signals the difficulty of annotating certain units.

We find a handful of patterns that make these items semantically nuanced, for example the presence of perception verbs (e.g., saw, heard). These verbs can notoriously take on different readings (Dik and Hengeveld, 1991), and they render a clause suitable to being an event from the perspective of an external observer, but also a subjective experience if one considers how a character has been affected by that event. Similarly, short exclamations (e.g., "*Good Lord!*") could be seen as descriptions of observable acts (e.g., that of speaking) or as indicative of a character's stance. The same holds for clauses that denote external behaviors while symbolizing internal modes of being (e.g., crying).

The ambiguity of some items might have stemmed from layered information as well. That is the case, for instance, when clauses include both a personal judgment and a dialogue marker (e.g., "*"You judge very properly," said Mr. Bennet.*"), such that the resulting label depends on what part of the text the readers focused on. Of this type are also clauses including manner adverbs, where objective events are qualified from a personal perspective ("*He was resolutely silent*").

Lastly, errors in the text preprocessing might have played a role, making the annotation more difficult for incorrect clause splits.

7. Story Structures

Does our schema reveal meaningful story patterns? With our second research question, we aim at understanding if the narrative content of novels corresponds to their description in terms of subjective experiences, events, and contextual information. In other words, we study if our theory-driven annotations reflect any interesting feature of the books in CLAUSE-ATLAS.

We answer this question at different scales of resolution, to account for the idea that stories are comprehended at multiple interacting granularities (cf. Section 3): first, at a coarse-grained level, looking at a book as a whole; second, at a chunk level, where a book is an ordered sequence of blocks of text; and lastly, at the level of clauses. In all of them, we use the labels generated by P1 to see how much our most reliable worker captured qualitative aspects of stories, but we could do the same using any other annotator: our objective is not to assess if the perceived narrative structure of a book is the correct one, but how a reader recognizes relative changes between one type of information to another, and whether such an interpretation (be it automatic or human) can be made sense of.

Book Level. As already seen in Figure 4, the majority of clauses in all books is about the internal life of characters. At the same time, clear differences between novels emerge from the relative proportion of the labels. In Alice's Adventures in Wonderland and The Adventures of Pinocchio, contextual information is comparatively lower than events, as opposed to Peter Pan and The Great Gatsby, where such labels are detected to the same extent. A possible explanation for this outcome is that the former two books follow the exploration of the protagonists in a world that is new to them: the narration provides less information to prompt one's understanding of objective states of affairs, and more to put oneself in the shoes of the characters (e.g., "*but she could not even get her head through the doorway;*" could be contextual information, but in its specific narrative frame, it conveys the physical changes that Alice undergoes in Wonderland).

Pride and Prejudice and Frankenstein, Victorian novels published in 1813 and 1818, have approximately the same histogram.[14] Both develop around a lower amount of events compared to the children's books. Consider, e.g., Pride and Prejudice against Peter Pan. Their difference might be due to the former elaborating more on the characters' development and interactions. Its focus moves from a social setting to another (balls, family estates, intimate discussions), reflecting in more contextual information to clarify the characters' relations and

[13] P1 reaches agreement close to humans also in regards to experiencers: the average Cohen's κ (1960) between pairs of human annotators is 51.3, and κ=50.3 when computed between them and P1.

[14] The three label distributions are not statistically different: p-value$>$0.3 with Chi-square test (Pearson, 1900).

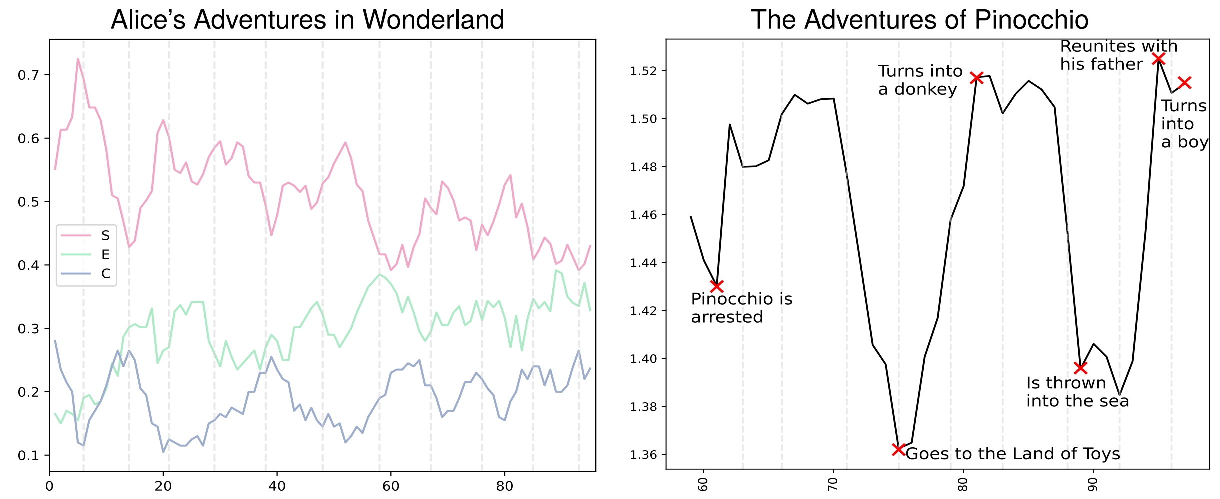

Figure 5: Left: Label distribution in Alice's Adventures in Wonderland (y-axis: relative frequencies), with a descending trend of subjective experiences (S), gradually reaching an equilibrium with the other labels (E: event; C: cont.info.). Right: Label entropy (y-axis) in the last 9 chapters of Pinocchio. In both plots, the x-axis reports book chunk numbers (out of 100), grey lines represent chapter boundaries. Both statistics were computed using a moving average with a window size of 6 chunks to remove high-frequency noise.

status. Multiple characters appear in Peter Pan as well. However, in this case the central themes are more dynamic (the travel in Neverland, the antagonism with Captain Hook), resulting in sequences dense of adventures and battles (i.e., events).

Chunk Level. We split the books into 100 chunks with approximately the same number of clauses, and we study how the labels unfold throughout a story. Figure 5 (left) shows the frequency curves of the three labels using Alice's Adventures in Wonderland as a use case. Subjective experiences, which initially dominate the book, slowly decrease: as the story begins, the reader learns to be traveling close to Alice, with the world being described through her eyes, via multiple internal monologues; later, a trade-off between different types of information takes place, suggesting that the reader gets acquainted with her point of view, discerning more and more objective facts.

Interesting insights also emerge by looking at the entropy of the labels as a summary statistics. Figure 5 (right) reports this information for Pinocchio, zooming in the last 9 chapters. Both particularly high entropy scores (corresponding to spikes of one or another label) and particularly low entropy scores (corresponding to more equilibrium between the labels) appear at crucial narrative points of the book (i.e., much is happening inside the characters or in their surrounding), such as disruptive events that put Pinocchio into danger, or that resolve his misfortunes.

Clause Level. Our annotation schema turns a book into a temporal sequence of labels, which allows us to examine transitions between different types of clauses. For this last analysis, we make a simplifying second-order Markovian assumption and study if the probability of a label is a function of preceding one or two labels (e.g., p(event | event, subj.exp.)). We find common patterns across the books: the narrative discourse more likely transits from events to subjective experiences than the other way around (i.e., the narrator lingers on the effects that events have on the characters). Further, labels tend to group together (e.g., p(event | event, event) > (e.g., p(event | cont.info., subj.exp.)). This is especially the case for children's books, which appear to have less unstable narrative structures in respect to switches between types of information.

8. Discussion and Conclusion

In this paper, we propose a methodology to create data for literary studies in NLP with a new type of annotator and a new annotation schema. The reported experimental results hint at the possibility to efficiently analyze literary texts at a large scale using generative LLMs. In particular, `gpt-3.5-turbo` matches natural readers in terms of story understanding based on the selected narrative theory. Its outputs are expressive enough to describe the structure and dynamical progression of novels, by moving up (chunks, chapters) and down (clauses) the natural organization of a book. We conclude that the chosen annotation schema can be applied to more novels than the six currently present in CLAUSE-ATLAS.

Our annotation pipeline maintains the same channel for structured extraction (implemented via function calling) across different input prompts. Although the variation of prompts has an effect on the heterogeneity of the labeling choices, the agreement profiles of our automatic annotators are analogous to the humans'. This outcome indicates that prompts (can) represent separate annotators, much in the way that humans display different annotation behaviors. CLAUSE-ATLAS could thus be used in a *data perspectivism* research framework (Cabitza et al., 2023), with the idea that disagreements do not necessarily need resolution but often point at different understandings of language. Studying if LLMs can be prompted to take on specific literary perspectives could be an addition to our findings.

Indeed, as an exploratory analysis, our work left open various questions. For example, ChatGPT's performance is still to be evaluated using more recent releases and various model temperatures; additionally, given the context-dependence of our labeling task, it would be important to grant LLMs access to inter-paragraph dependencies. The theory we follow is ripe for exploration as well. If extended to more data from individual authors, it could facilitate the performance of robust stylistic analyses. Notably, it holds potential to bridge research on events and characters (which we collected but did not focus on): investigating the link between the two (e.g., what does a character experience when something happens?) could boost our understanding of human story comprehension, and the attempt to endow computational systems with the same ability.

Acknowledgements

This research was founded by the CLARIAH project. We also thank our annotators, and Annika Kniele, Jaap Kruijt, and Levi Remijnse, for their participation and useful feedback.

9. Bibliographical References

Jon-K Adams. 1989. Causality and narrative.

Apoorv Agarwal, Augusto Corvalan, Jacob Jensen, and Owen Rambow. 2012. Social network analysis of alice in wonderland. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 88–96, Montréal, Canada. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

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

Ruth A. Berman. 1997. Narrative theory and narrative development: The labovian impact. *Journal of Narrative and Life History*, 7(1-4):235–244.

Megan Biesele. 1986. How hunter-gatherers' stories "make sense": Semantics and adaptation. *Cultural Anthropology*, 1(2):157–170.

Thomas Bögel, Jannik Strötgen, and Michael Gertz. 2014. Computational narratology: Extracting tense clusters from narrative texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 950–955, Reykjavik, Iceland. European Language Resources Association (ELRA).

Brian Boyd. 2009. *On the Origin of Stories: Evolution, Cognition, and Fiction*. Harvard University Press.

Brian Boyd. 2018. The evolution of stories: From mimesis to language, from fact to fiction. *Wiley Interdisciplinary Reviews: Cognitive Science*, 9(1):1–16.

Ryan L. Boyd. 2017. *Psychological Text Analysis in the Digital Humanities*, pages 161–189. Springer International Publishing.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cleanth Brooks. 1951. The formalist critics. *The Kenyon Review*, 13(1):72–81.

Joseph Bullard and Cecilia Ovesdotter Alm. 2014. Computational analysis to explore authors' depiction of characters. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 11–16, Gothenburg, Sweden. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

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

Mariona Coll Ardanuy and Caroline Sporleder. 2014. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 31–39, Gothenburg, Sweden. Association for Computational Linguistics.

Gregory Currie. 2009. Narrative and the Psychology of Character. *The Journal of Aesthetics and Art Criticism*, 67(1):61–71.

Jerome De Groot. 2009. *The historical novel*. Routledge.

Simon C Dik and Kees Hengeveld. 1991. The hierarchical structure of the clause and the typology of perception-verb complements.

Anca Dumitrache. 2019. *Truth in Disagreement: Crowdsourcing Labeled Data for Natural Language Processing*. Ph.D. thesis, Vrije Universiteit Amsterdam.

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam, Gintarė Grigonytė, and Sofia Gustafson Capková. 2018. Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stanley Fish. 1970. Literature in the reader: Affective stylistics. *New literary history*, 2(1):123–162.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal. Association for Computational Linguistics.

Monika Fludernik. 2005. Histories of narrative theory (ii): From structuralism to the present. *A companion to narrative theory*, pages 36–59.

Thomas C. Foster. 2003. *How to read literature like a professor: A lively and entertaining guide to reading between the lines*. Quill New York.

Northrop Frye. 1951. The archetypes of literature. *The Kenyon Review*, 13(1):92–110.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA. Association for Computational Linguistics.

Ina Habermann and Nikolaus Kuhn. 2011. Sustainable fictions – geographical, literary and cultural intersections in J.R.R. Tolkien's The Lord of the Rings. *The Cartographic Journal*, 48(4):263–273.

Adam Hammond, Julian Brooke, and Graeme Hirst. 2013. A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8, Atlanta, Georgia. Association for Computational Linguistics.

David Herman. 1997. Scripts, sequences, and stories: Elements of a postclassical narratology. *PMLA*, 112(5):1046–1059.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Using chatgpt to fight misinformation: Chatgpt nails 72% of 12,000 verified claims. *PsyArXiv*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Forrest L. Ingram. 1971. *Representative Short Story Cycles of the Twentieth Century*. De Gruyter Mouton, Berlin, Boston.

Wolfgang Iser. 1979. The act of reading: A theory of aesthetic response. *Journal of Aesthetics and Art Criticism*, 38(1):88–91.

Allen Kim, Charuta Pethe, and Steve Skiena. 2020. What time is it? temporal analysis of novels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9076–9086, Online. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge University Press.

Aleksandra Konovalova and Antonio Toral. 2022. Man vs. machine: Extracting character networks from human and machine translations. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 75–82, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taja Kuzman, Igor Mozetic, and Nikola Ljubešic. 2023. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *ArXiv, abs/2303.03953*.

William Labov. 1972. *Language in the inner city*, volume 3. Philadelphia: University of Pennsylvania Press.

William Labov. 2013. *The language of life and death: The transformation of experience in oral narrative*. Cambridge University Press.

William Labov and Joshua Waletzky. 1997. Narrative analysis: oral versions of personal experience. *Journal of Narrative and Life History*, 7(1–4):3–38.

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Dane Malenfant Margaret Meehan and Andrew Piper. 2022. Causality mining in fiction. In *Proceedings of Text2Story-Fifth Workshop on Narrative Extraction From Texts*, volume 3117, pages 25–34.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Yuji Oka and Kazuaki Ando. 2020. Extraction of novel character information from synopses of fantasy novels in Japanese using sequence labeling. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 505–513, Hanoi, Vietnam. Association for Computational Linguistics.

Jessica Ouyang and Kathy McKeown. 2014. Towards automatic detection of narrative structure. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4624–4631, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Federico Pianzola. 2018. *Looking at Narrative as a Complex System: The Proteus Principle*, pages 101–122. Springer International Publishing, Cham.

Andrew Piper. 2022. The conlit dataset of contemporary literature. *Journal of Open Humanities Data*, 8.

Andrew Piper, 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*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniele Pizzolli and Carlo Strapparava. 2019. Personality traits recognition in literary texts. In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.

Gerald Prince. 2003. *A dictionary of narratology*. Lincoln, NE, USA: University of Nebraska Press.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 979–988, Uppsala, Sweden. Association for Computational Linguistics.

Michael V. Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Nils Reiter. 2015. Towards annotating narrative segments. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 34–38, Beijing, China. Association for Computational Linguistics.

James S Romm. 1994. *The edges of the earth in ancient thought: geography, exploration, and fiction*. Princeton University Press.

Antonio Roque. 2012. Towards a computational approach to literary text analysis. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 97–104, Montréal, Canada. Association for Computational Linguistics.

Belen Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 78–86, Online. Association for Computational Linguistics.

Michelle Scalise Sugiyama. 2011. The forager oral tradition and the evolution of prolonged juvenility. *Frontiers in Psychology*, 2.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Sara Stymne and Carin Östman. 2020. SLäNDa: An annotated corpus of narrative and dialogue in Swedish literary fiction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 826–834, Marseille, France. European Language Resources Association.

Michelle Scalise Sugiyama. 2005. *Reverse-Engineering Narrative: Evidence of Special Design*, pages 177–196. Northwestern University Press.

Peiqi Sui, Lin Wang, Sil Hamilton, Thorsten Ries, Kelvin Wong, and Stephen Wong. 2023. Mrs. dalloway said she would segment the chapters herself. In *Proceedings of the The 5th Workshop on Narrative Understanding*, pages 92–105, Toronto, Canada. Association for Computational Linguistics.

Stith Thompson. 1955. *Motif-Index of Folk-Literature, Volume 4: A Classification of Narrative Elements in Folk Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*, volume 4. Indiana University Press.

Alexey Tikhonov, Igor Samenko, and Ivan Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 32–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lynne Tirrell. 1990. Storytelling and moral agency. *The Journal of Aesthetics and Art Criticism*, 48(2):115–126.

Paulina Toro Isaza, 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)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 184–189, Portorož, Slovenia. European Language Resources Association (ELRA).

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

S. S. Van Dine. 1928. Twenty rules for writing detective fiction. *The Art of the Mystery Story*.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In *CHR*, pages 333–345.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Janyce M. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Polly W. Wiessner. 2014. Embers of society: Firelight talk among the ju/'hoansi bushmen. *Proceedings of the National Academy of Sciences*, 111(39):14027–14035.

W. Victor H. Yarlott and Mark A. Finlayson. 2016. Proppml: A complete annotation scheme for proppian morphologies. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023a. Assessing the potential of ai-assisted pragmatic annotation: The case of apologies. *arXiv preprint arXiv:2305.08339*.

Juntao Yu, Silviu Paun, Maris Camilleri, Paloma Garcia, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2023b. Aggregating crowdsourced and automatic judgments to scale up a corpus of anaphoric reference for fiction and Wikipedia texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 767–781, Dubrovnik, Croatia. Association for Computational Linguistics.

Albin Zehe, Leonard Konle, Lea Katharina 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: Main Volume*, pages 3167–3177, Online. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

A. Inspection of Clause Quality

Tools for clause extraction are not easily accessible or ready-to-use (e.g., constituency parsers, typically designed for broader purposes, need customization to handle specific cases, like nested clauses). As a straightforward solution, we obtained our annotation units by preprocessing the six books with the following steps:

1. we used regular expressions to match and remove noise from the book downloaded from Project Gutenberg (e.g., notes, image placeholders, chapter subtitles);
2. we divided the books into paragraphs, and the paragraphs into sentences (with the python library spacy);
3. we orchestrated the LLM calls, using the OpenAI chat completion API with `gpt-3.5-turbo`, enforcing a json structured extraction to segment each sentence into clauses.

The textual strings found this way are not all coherent from a linguistic standpoint. Some of them correctly correspond to individual clauses; others are long text chunks spanning multiple smaller clauses. This observation requires us to clarify the quality of the textual segments in CLAUSE-ATLAS. Narrowing the analysis to a portion of the corpus, we evaluated the quality of 120 clauses from four paragraphs randomly sampled from each book.

This annotation was carried out by three PhD students of NLP (with a Master's degree in linguistics) tasked to decide if a given clause represented a "good" clause or, on the contrary, a chunking mistake. The annotators did not receive in-depth training based on general linguistics aspects (e.g., on the differences between types of clauses). Instead, they were given the minimal definition used to prompt the LLM. They considered a "good" clause one that is a semantically and grammatically self-contained substructure of a sentence, as an appropriate unit for our purpose of narrative annotation. Most of the clauses passed the inspection. For 95.8% of the items, at least two annotators agreed that the chunking output was correct. Further, the three annotators unanimously indicated that the LLM made no mistake on 74% of the items.

While satisfied with these scores, we realize that literary analyses are strongly dependent on this preprocessing passage. For the future, it would be useful to conduct a more comprehensive evaluation of the text segmentation capabilities of generative LLMs under different theoretical conditions (e.g., to apply other narrative schemas to the data), and focusing on different linguistic resolutions (e.g., syntactic constituents).