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

Enrica Troiano and Piek Vossen
CLTL, Vrije Universiteit Amsterdam
e.troiano@vu.nl

VU VRIJE UNIVERSITEIT AMSTERDAM
Clariah

We create an "atlas" of 41,715 clauses from six English novels, automatically annotated with narrative information that captures how stories unfold.

Framework: Computational Story Understanding

Ideal goal:
Structural level: abstracting over words, sentences, paragraphs, ...
Conceptual level: extracting actions, characters, time, ...

First step:
Structural level: focus on tensed clauses
Conceptual level: distinguishing types of narrative information about the characters or the narrative world:

- Events: all that happens outside the characters
- Subjective Experiences: what happens inside the characters, (thoughts, perceptions, emotions, memories)
- Contextual Information: extra details about the story

Achilles withdrew from the battle, → Event
filled with anger and resentment → Subj. Exp.
for the conflict with Agamemnon. → Cont. Information

book	chapter_id	paragraph_id	clause_number	text	prompt_one
Alice's Adventures in Wonderland	1	1	0	Alice was beginning to get very tired of…	C
Alice's Adventures in Wonderland	1	1	1	once or twice she had peeped into the book…	S
Alice's Adventures in Wonderland	1	1	2	"and what is the use of a book," thought…	S
Alice's Adventures in Wonderland	1	2	3	So she was considering in her…	S
Alice's Adventures in Wonderland	1	2	4	whether the pleasure of making a daisy-…	S
Alice's Adventures in Wonderland	1	2	5	when suddenly a White Rabbit with pink eye…	E

Events, **S**ubjective experiences, and **C**ontextual information annotate entire books in CLAUSE-ATLAS (**prompt_one** is one annotator).

ALICE'S ADVENTURES IN WONDERLAND
PETER PAN
THE ADVENTURES OF PINOCCHIO
FRANKENSTEIN
PRIDE AND PREJUDICE
THE GREAT GATSBY

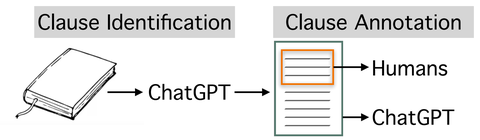
Challenge: Heavy Human Annotation Labour

Clause Identification Clause Annotation

Is ChatGPT a good literary annotator?

It can handle both tasks, but its narrative understanding must be probed.

Materials and Methods



LLM-informed pipeline, based on `gpt-3.5-turbo` with 16k tokens context (accessed via the official OpenAI's API).

1. Clause identification
Find all clause boundaries in a given paragraph
(Clauses pass manual quality inspection)

2. Clause annotation
Task: one clause → one label
Result: 6 annotation layers

Automatic Setup: annotate all books

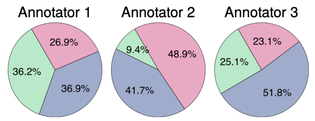
- Classifiers implemented through ChatGPT prompting strategies
- 3 different input channels (prompts P1, P2, P3)
- fixed output channel with function calling

Human Setup: annotate first chapter of The Great Gatsby, Alice's Adventures in Wonderland, The Adventures of Pinocchio

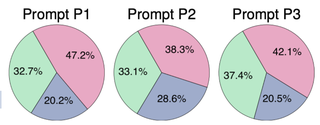
- 3 Master's students with background in literary studies
- guidelines = most detailed prompt (P3)

Inter-Annotator Agreement

Comparison on the data sample annotated by both.



Events
Subj.Experience
Cont.Information



Greater agreement in the automatic setup

	Fleiss' κ		
	Humans	ChatGPT	Both
First Chapter	.32	**.58**	.33
All	–	.57	–

Results:

- P1, P2, P3 converge more often than not
- Perfect agreement on 60% items
- Human–classifier pairs agree on 49% items (≈ humans only)
- People agree more where prompts spurred perfect agreement
- Classifiers and humans disagree with the same linguistic patterns

Disagreement Patterns	Example
Negations	*but Slightly had the sense not to see her*
Manner Adverbs	*He was resolutely silent*
Perception Verbs	*... you saw it was a fairy*
Segmentation Errors	*"You judge very properly", said Mr. Bennet.*

ChatGPT agrees with humans to a comparable extent as humans agree among each other.

Analysis of Novels

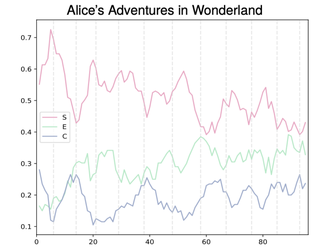
Clause Level

p(Event | Event, Event) > p(Event | Cont.Info., Subj.Exp.)

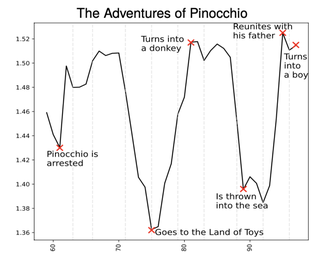
E.g., Pinocchio has "slower" narrative switches than Pride and Prejudice (in terms of types of information)

Within Books

Subj. Experiences decrease as the story proceeds



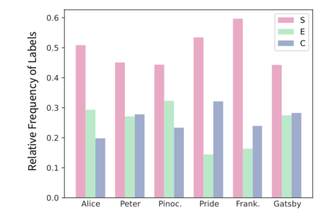
Peaks of entropy correspond to key narrative points



Across Books

Similar histograms:

- C < E (Alice, Pinocchio)
- C ≈ E (Peter, Gatsby)
- high C (Pride, Frank.)



ChatGPT's understanding of events, subjective experiences and contextual information capture story developments.

Download CLAUSE-ATLAS