# Distinguishing Fictional Voices: a Study of Authorship Verification Models for Quotation Attribution

Gaspard Michel, Elena V. Epure, Romain Hennequin, Christophe Cerisara

# Distinguishing Fictional Voices: a Study of Authorship Verification Models for Quotation Attribution

**Gaspard Michel**[†*] gmichel@deezer.com
**Elena V. Epure**[†] eepure@deezer.com
**Romain Hennequin**[†] rhennequin@deezer.com
**Christophe Cerisara**[*] christophe.cerisara@loria.fr

[†] Deezer Research, Paris, France
[*] Loria, Nancy, France



## Abstract

Recent approaches to automatically detect the speaker of an utterance of direct speech often disregard general information about characters in favor of local information found in the context, such as surrounding mentions of entities. In this work, we explore stylistic representations of characters built by encoding their quotes with off-the-shelf pretrained Authorship Verification models in a large corpus of English novels (the Project Dialogism Novel Corpus). Results suggest that the combination of stylistic and topical information captured in some of these models accurately distinguish characters among each other, but does not necessarily improve over semantic-only models when attributing quotes. However, these results vary across novels and more investigation of stylometric models particularly tailored for literary texts and the study of characters should be conducted.

## 1 Introduction

In prose fiction, entire universes come to life. Different techniques are employed by authors to create engaging narratives and use a combination of narrator and character words to build the atmosphere and unveil the story. Characters in the fictional world reveal aspects of their personalities through dialogues. In Bakhtin's idea of *polyphony* (Bakhtin, 1984), characters participate in dialogues in their own voice, according to their own ideas about themselves and the fictional world. Automatically identifying parts of dialogues and attributing them to the character that utters them is central to many studies of large literary corpora (Elson et al., 2010; Muzny et al., 2017a; Sims and Bamman, 2020)

The detection of direct-speech has been widely performed for English literature, and simple regular expression systems achieve almost perfect performances on well-formatted texts. Attributing characters to quotes is more challenging and often requires solving multiple tasks: quotation identification, character identification and speaker attribution (Muzny et al., 2017b; Vishnubhotla et al., 2023). A speaker is attributed to a quote by training a separate model to find the nearest relevant *entity mention*, which is then linked to a *canonical character* with coreference resolution models. Figure 1 summarizes this process. Although many approaches have been explored in this direction, there is still room for improvement

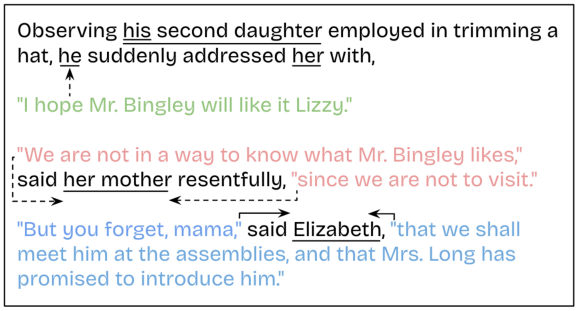


Figure 1: Example of quotation attribution on an excerpt of *Pride and Prejudice* by Jane Austen (1813). Underlined text are identified mentions, and arrows link quotes to their relevant entity mention (solid arrows are explicit references and dashed arrows are anaphoric references). In a separate step, coreference resolution is used to link entity mentions to their canonical character.

Using a recently proposed corpora of English novels annotated with speakers, our current work first investigates to which extent *voices* of characters in novels are distinguishable using authorship verification approaches applied to character utterances. Then, we analyze quotes of characters in this large corpus and evaluate to which extent character-related features encoded by pretrained authorship verification models contain a predictive signal for *quotation attribution*. Our intuition is that character-level information (such as style, preferences in topic, persona) might be used in addition to contextual information to improve quotation at-

tribution models. Prior stylometric studies have shown that canonical drama authors are able to create memorable characters with distinguishable voices (Vishnubhotla et al., 2019; Šeļa et al., 2023). Nonetheless, the stylometric analysis of characters in novels remains scarce, mainly due to the lack of available corpora annotated with speakers. To the best of our knowledge, exploring this type of character representations for quotation attribution has not been done before.

Consequently, with this work, we make the following contributions:

1. We investigate recent neural authorship verification models for the study of characters in novels and benchmark them on their ability to attribute authorship for distinguishing character voices in a large corpus.

2. Framing quote attribution as an authorship verification task, we are the first work to evaluate the usefulness of stylometric character representations encoded by off-the-shelf authorship verification models to attribute quotes to characters.

Results suggest that most characters in the PDNC corpus own distinct voices, and that they are best distinguished by models that encode both semantic and stylistic information. Semantic-only models, however, seem to be better at attributing quotes than models that encode style. Besides, representing characters with the quotes they uttered in a single chapter appear to contain a predictive signal for attributing quotes in other chapters, but this varies per novel. Finally, our results suggest that there are semantic variations between *explicit* quotes (i.e. quotes where the relevant gold mention is a named mention of the speaker) and other type of quotes (*anaphoric* and *implicit* quotes, introduced in Section 3), and that including stylistic information alleviates the impact of these semantic shifts when distinguishing characters voices based on *explicit* quotes only.

## 2 Related Work

### 2.1 Quotation Attribution

Quotation attribution models in novels often assume given utterances of direct speech. Elson and McKeown (2010) introduce the CQSC corpus and attribute automatically extracted quotes to named entities ("Elizabeth") and nominals ("her daughter") with a mention ranking model. Instead, He et al. (2013) attribute quotes directly to *speakers* with a supervised ranking system using features such as speaker alternation patterns and character-level features (He et al., 2010). The deterministic sieve-based model of Muzny et al. (2017b) regards quotation attribution as a two-step process: quote-*mention* linking and mention-*speaker* linking.

The NLP pipeline dedicated to books, BookNLP[1], went a step further by replacing the deterministic sieves with fine-tuned language models. Vishnubhotla et al. (2022) introduce the largest-to-date corpus of quotation attribution, PDNC, and show a similar accuracy score of around 63% for both BookNLP and the sieve-based model. However, better results were obtained later by fine-tuning BookNLP on PDNC (Vishnubhotla et al., 2023). Although these works are considered state-of-the-art in quotation attribution, they inherently lack character-level information in the mention-speaker linking step.

### 2.2 Character Representations

Most works focus on creating distributed *embeddings* that encode the persona of characters (i.e. characters with similar properties such as gender, job, relationships should have similar persona-based representations). Bamman et al. (2014) propose a Bayesian model that infers latent character personas as a distribution over various dependency relations. Brahman et al. (2021) introduce LiSCU, a dataset containing literary texts along with their summaries and descriptions of characters participating in the narrative. They train a language model to generate accurate descriptions of characters, showing that the model has a complex understanding of personas. Inoue et al. (2022) propose to represent characters in novels using a graph-based character network and positional embeddings. The character network contains book-level and authorial information, and captures the attributes of characters, while positional embeddings encode the dynamics of character activity throughout the narrative.

In this work, we rather focus on *what characters say* and *how they say it*, building stylometric representations of characters with off-the-shelf pretrained authorship verification models.

### 2.3 Authorship Verification

Authorship verification aims to predict whether two texts have been written by the same author. Re-

[1] https://github.com/booknlp/booknlp

cently, these models have employed a contrastive learning framework to build a representation space where works written by the same author are close together while being distant to texts written by other authors. Evaluation is made by building disjoint sets of *queries* and *targets*. Queries are pieces of text written by an author, and targets are other texts written by the same author and other authors. Based on a similarity measure such as cosine similarity, a ranking distribution is created by scoring a query against all targets. Area Under the Receiver Operating Characteristic Curve (AUC) is often used to evaluate if this distribution gives a high rank to the correct target.

Recent advances exploit language models to distinguish hundred of thousands of authors. Rivera-Soto et al. (2021) fine-tune SentenceBERT (Reimers and Gurevych, 2019) using thousands of Reddit users, Amazon reviews and Fanfiction stories. Their model, LUAR, uses both stylistic and content information (such as topical preferences) to distinguish between authors. Similarly, Wegmann et al. (2022) fine-tune RoBERTa (Liu et al., 2019) on posts of thousands of Reddit users. By controlling for content in the creation of their training data, they ensure that their model, STEL, mostly encodes stylistic information.

Although both models perform well on their respective authorship verification tasks, they are blackbox models that do not offer an interpretation of aspects of style captured in their representations. Wegmann et al. (2022) present a clustering analysis of learned representations of Reddit posts and showed that STEL mostly captures variations of punctuation, casing and contraction spelling. These stylistic variations do not apply to quotes in novel, we thus expect STEL to struggle to transfer from Reddit to our domain. Rivera-Soto et al. (2021) do not study aspects of style captured in LUAR's representations, but offer an interpretation of the model's performance based on the domain it was trained on. Particularly, they show that LUAR is prone to overfit to the training domain style features. While training on Reddit data, they conclude that the model rely less on topical diversity to distinguish among authors, which can be favorable to distinguish novel characters that usually speak in a wide range of topics.

### 2.4 Stylometric Analysis of Characters

Stylometric analysis of literary characters has been mostly focused on drama characters because of existing large annotated corpus. Most works focus on the style of character, aiming at capturing syntactic, lexical and phonological variations that can occur when they are quoted. Vishnubhotla et al. (2019) propose to study the distinctiveness of character stylistic and topical patterns with text classification. Šeļa et al. (2023) propose a measure of distinctiveness based on character 3-grams, which they apply to a large number of drama characters. They show that it is able to capture interesting aspects of stylistics such as phonological differences, accents and dialects, as well as topical and lexical differences. To the best of our knowledge, Dinu and Uban (2017) is the only work that focuses on novel characters. Their supervised bag-of-word classification model was able to accurately classify some characters, but fell short on the main character of the epistolary novel "Liaisons Dangereuse".

Other related works leverage dialogues in movie scripts to build character representations. Azab et al. (2019) train a Word2Vec model where the context window consists of the surrounding speaker identities as well as the current speaker utterance. Similarly, Li et al. (2023) encode all utterances of a script with a pre-trained language model, and extract representations by pooling all encoded quotes of a character together. A contrastive learning objective is used to create a fine-grained representation space where characters are well separated. Although the methods presented above are quite similar to the way we build character representations, authors did not release the code publicly at the time of writing, precluding comparison in our experiments.

In this work, we analyze novel characters at a larger scale using the PDNC corpus containing 28 English novels. Instead of employing classification accuracy as a measure of character distinctiveness, we frame the task as an authorship verification problem to evaluate to what extent characters voice can be distinguished. We are also the first to evaluate if these character-level features contain a predictive signal to attribute unseen quotes to the right speaker.

## 3 Experimental Setup

Our goal in this work is to investigate if fictional voices of literary characters in novels are distin-

guishable from a stylistic point-of-view. We also want to know if a partial signal of a character's voice derived from its *explicit* quotes (i.e. the gold *mention* linked to the quote is any named mention such as "Elizabeth") is a good proxy for its overall voice. Explicit quotes are straightforward to attribute to characters since they are linked to a named mention, which can then be linked to canonical characters (*e.g.* with coreference resolution or name clustering) more easily than when dealing with pronominal mentions (Muzny et al., 2017b). We hypothesize that if we can construct representative character embeddings based on explicit quotes only, then these representations can in turn enhance quotation attribution solutions to detect the speaker of other type of quotes. Other types of quotes include *anaphoric* quotes (i.e the gold *mention* is a pronoun or noun phrase) and *implicit* quotes (often happens during a conversation, when no *mention* is linked to the quote but the speaker can be inferred from the context). Finally, using the same set of character representations, we want to evaluate to which extent they contain information to attribute quotes that were not used to build the representations.

To evaluate these representations, we formulate the task as the authorship verification task: given a corpus of quotes from character A (the *query*), a corpus of other quotes from character A and similar corpora for other characters in a given novel (the *targets*) and a similarity measure, we evaluate the ability of pretrained models to predict if the targets have been written by character A or not. AUC is used to assess models' performances, as it accounts for how well models can rank predictions, without concerns of threshold values (Tyo et al., 2022). We chose to frame the task as an authorship verification problem rather than closed-set authorship attribution because the number of targets (*i.e.* number of candidate speakers) vary for each query, which is further described in Section 3.4

We first present how character representations are derived from pretrained models, and then describe how we evaluate the capacity of these representations to answer the above questions. We publicly release our code for further research[2].

### 3.1 Building Character Representations

Transformer-based models are widely used to encode textual information. To build character representations, we leverage various publicly available pretrained models (PM) trained on different tasks as quote encoders. For each novel, we assume that we have access to all utterances of direct speech $Q = \{q_1, \ldots, q_n\}$ as well as each character in the novel $C = \{c_1, \ldots, c_m\}$. Let $g : Q \mapsto C$ be a function that assigns a quote $q_i$ to its speaker $c$ such that $g(q_i) = c$ implies that character $c$ is the speaker of the quote $q_i$. To build the representation of a character $c$ in a given subset of quotes $\tilde{Q} \subset Q$, we first extract all quotes of character $c$ in the subset: $\tilde{Q}_c = \{q_i : q_i \in \tilde{Q}, g(q_i) = c\}$. A quote representation is obtained by encoding each quote with a pretrained model, denoted as $\text{PM}_\theta$:

$$\mathbf{h}_{q_i} = \text{PM}_\theta(q_i)$$

We then derive an embedding of character $c$ in the subset $\tilde{Q}$ by pooling all embeddings of quotes spoken by $c$ in $\tilde{Q}$:

$$\mathbf{H}_{\tilde{Q}_c} = \text{POOL}(\{\mathbf{h}_{q_i} : q_i \in \tilde{Q}_c\})$$

In our experiments, the POOL function is the average of all quote representations, except for the LUAR model that uses an attention-based POOL function with attention weights trained to focus on the relevant texts of an author. By pooling over the subset of quotes of a character, we expect the resulting representation to contain general information of *what a character say* and/or *how he says it*, depending on the PM used. Compared to some of the previous approaches to character representations presented in section 2.2, we do not use any contextual information (surrounding passages of narrative text, sequence of speaker turn, or surrounding quotes) so that that the representations focus mainly on stylistic and/or content information. We conduct different experiments by varying the construction of the subset $\tilde{Q}$.

**Chapterwise**: we extract all quotes of a character in a given chapter $T$ to build its query representation. The targets are created by using quotes contained in the held-out chapters.

**Explicit**: we only extract *explicit* quotes of a character in a given chapter $T$ with similar targets as in the chapterwise experiment. We thus build representations for a character with quotes that are linked to a named mention of the character. This experiment is designed such that we can quantify the amount of information lost compared to the chapterwise experiment that uses all types of quotes. It

[2] https://github.com/deezer/quote_AV

might happen that some characters are not explicitly quoted in chapter $T$. In this case, we do not build representations for these characters.

**Reading Order**: we use the first $n$ quotes of a character in the first half of novels (segmented by chapter) as a basis of its query representation. Targets are built using quotes in the remaining half. With this experiment, we want to see the impact of increasing the amount of available character information on the capacity of models to distinguish their voices.

### 3.2 Data

We use the PDNC dataset[3] (Vishnubhotla et al., 2022), containing annotations of speakers at the quote level for 28 English novels written by 21 authors and published between the 19th and early 20th century. This dataset consists of mostly literary novels, and a few children, crime and science-fiction novels. Characters in each novel are labelled with *minor*, *intermediate* and *major* roles, depending on the total number of quotes they uttered. We only focus on *intermediate* and *major* characters that uttered at least 10 and 100 quotes respectively, and discarded *minor* characters that participate less in the narrative. Quotes are often subject to *incises*, where a narrative segment giving indication on who and how the quotes is being said is inserted within the quote (*e.g.* "said her mother resentfully", third paragraph in Figure 1). In this case, we use the full text of the quote, discarding the incise, as a single character's utterance.

We build character embeddings using the methodology explained in Section 3.1 that we further use to derive sets of queries and targets. For a character $c$ and its quote subset $\tilde{Q}_c$, the associated query is the character representation built from the subset, $\mathbf{H}_{\tilde{Q}_c}$. Using the held-out subset, $O$, the associated set of targets are embeddings of every character that utters quotes in $O$: $\{\mathbf{H}_{O_{c'}}: \quad c' \in C\}$. We only construct queries for characters that utter at least 5 quotes in $\tilde{Q}_c$ to mitigate the amount of uninformative queries. We chose to use 5 quotes based on preliminary results showing that some queries would have only 1 quote and that the resulting character representations were not really informative. Results of the reading order experiment presented in Section 4.3 further support this observation.

[3] https://github.com/Priya22/project-dialogism-novel-corpus

| | Chapterwise | Explicit |
|---|---|---|
| Total queries | 1606 | 562 |
| # Speakers | 11.1 (4.6) | 11.1 (4.6) |
| Activity (%) | 93 (10) | 53 (28) |
| Queries | 57.4 (29.3) | 21.6 (17.9) |
| Query length | 21.4 (11.5) | 10.5 (3.5) |
| Targets/query | | |
| Character | 11.0 (4.6) | 11.2 (4.7) |
| Quote | 1142 (600) | 1176 (597) |

Table 1: Summary statistics of our set of queries and targets on the PDNC corpus. Bottom part is averaged over novels with (standard deviation).

Table 1 summarizes the main statistics of the resulting data. For the chapterwise experiment, we derived 1606 queries on the entire corpus, but only 562 for the explicit experiment. Indeed, explicit quotes represent around 31% of the total number of quotes in the corpus, with large discrepancy across novels (the minimum is 10% and the maximum is 81%), thus leading to many characters that do not utter at least 5 explicit quotes in $\tilde{Q}_c$. As a result, the percentage of active characters (i.e characters that have at least one query) drops from 93% to 53% and, out of the 28 novels, we could not create queries for two novels because they did not contain enough explicit quotes to build representations (*The Gambler* by Fyodor Dostoevsky, 1887, and *The Sport of the Gods* by Paul Laurence Dunbar, 1902). Queries in the Chapterwise experiment also contain twice as many quotes on average than in the Explicit one. Nonetheless, the number of character targets and the number of quotes in target is roughly the same between the two experiments. We thus expect the task of distinguishing voices and attributing unseen quotes based on representation of explicit quotes only to be generally harder.

### 3.3 Models

We build representations with two pretrained authorship verification models: STEL and LUAR and introduce two baseline models: SentenceBERT (SBERT)[4] (Reimers and Gurevych, 2019) and a RoBERTa-based multi-label emotion classification model[5].

[4] We use `all-mpnet-base-v2`

[5] https://huggingface.co/SamLowe/roberta-base-go_emotions

#### 3.3.1 Baselines

The SBERT and Emotion models are referred to as baselines because their purpose is not to encode stylistic information of characters. SBERT is trained to recognize semantically similar sentences, hence encoding rich semantic textual information. We expect this semantic-only model to distinguish characters voices based on the content of their corpus of quotes, such as topical preferences. The Emotion model is a RoBERTa model fined-tuned to classify emotions in a multi-label setup (28 emotions), allowing predictions of multiple emotions conveyed at once in the same sentence. We use this model as a benchmark based on prior analysis we made, showing that some characters were conveying certain emotions more than others. Thus, our intuition was that it could be used as a discriminative feature of a character's voice. We use representations contained in the last RoBERTa transformer layer before the classification head when encoding quotes.

#### 3.3.2 Authorship Verification Models

Authorship verification models are trained to predict if two texts (or corpus of texts) have been written by the same author, enabling them to capture authorial style to some extent. STEL is a RoBERTa model fine-tuned on the Contrastive Authorship Verification (CAV) task with millions of Reddit users. In the CAV task, a model is asked to distinguish which from three pieces of texts (triplets) have been written by the same author. Using triplets that have similar topic, STEL is trained to distinguish between authors using stylistic information only. Although its base model, RoBERTa, encodes semantic to some extent, restraining training triplets to texts that essentially cover similar topic forces the model to focus on stylistic cues.

LUAR fine-tunes SentenceBERT to encode corpora of utterances. Unlike STEL, they do not force training examples to have similar topic, leading to a model that encodes both content and stylistic information in author representations. For the same author, different representations are built using distinct collections of documents written by the author. LUAR encodes stylistic information by being trained on the authorship verification task. Compared to STEL, we expect LUAR to build more robust character representations since it uses an attention mechanism that allows focus on texts with strong authorial signal.

For all models, we use a maximum sequence length of 64, truncating longer quotes (only 14% of the total number of quotes are longer than 64 tokens). We use publicly available versions of LUAR[6] and STEL[7].

### 3.4 Evaluation

#### 3.4.1 Character-Character

Given a query character representation built from a subset of quotes $\mathbf{H}_{\tilde{Q}_c}$, a similarity function $\phi$, and the held-out subset $O$, we evaluate the similarity of the query with the targets built from $O$. In this work, we use cosine similarity for the $\phi$ function. Ideally, we want a high similarity between a character query and the target linked to the same character, $\mathbf{H}_{O_c}$, and low similarity between the character query and other characters target $\mathbf{H}_{O_{c'}}$:

$$\phi(\mathbf{H}_{\tilde{Q}_c}, \mathbf{H}_{O_c}) > \phi(\mathbf{H}_{\tilde{Q}_c}, \mathbf{H}_{O_{c'}}),\ \forall\, c' \neq c \quad (1)$$

In practice, we evaluate the capacity of pretrained models to give a high rank to corresponding character target $\mathbf{H}_{O_c}$ using AUC. In this context, the AUC measures the probability that Equation 1 holds when randomly selecting a character $c'$ different than $c$. We chose AUC over standard authorship attribution metrics (such as macro-averaged accuracy) because of its ability to evaluate the output ranking distribution, $\{\phi(\mathbf{H}_{\tilde{Q}_c}, \mathbf{H}_{O'_c}), c' \in C\}$. Besides, unlike accuracy, AUC does not require a threshold value for predicting the speaker of a quote, which can be tricky when using cosine similarities. We refer to this evaluation as Character-Character (CC) as it measures how unique are characters voices.

#### 3.4.2 Character-Quotes

We now introduce how we evaluate the performances of such character representations at attributing quotes from the held-out subset. Similar query representations are used, but targets are replaced by quote representations (encoded by the same PM) rather than character representations. Let $q_i \in O_c$ be a target quote from character $c$ in the held-out subset $O$ and $q_j \in \bar{O}_c = \bigcup_{c' \neq c} O_{c'}$ be any quote spoken by a different character in $O$, we evaluate the following hypothesis:

$$\phi(\mathbf{H}_{\tilde{Q}_c}, \mathbf{h}_{q_i}) > \phi(\mathbf{H}_{\tilde{Q}_c}, \mathbf{h}_{q_j}),\ \forall\, q_i \in O_c,\ q_j \in \bar{O}_c \quad (2)$$

[6] https://huggingface.co/rrivera1849/LUAR-MUD

[7] https://huggingface.co/AnnaWegmann/Style-Embedding

| | CC | CQ |
|---|---|---|
| Semantics | 67.3 (11.6) | **55.1** (2.5) |
| STEL | 58.1 (8.3) | 52.8 (1.9) |
| Emotions | 56.0 (8.0) | 51.7 (1.5) |
| LUAR | **81.6** (6.2) | 53.6 (2.4) |

Table 2: AUC results of the **chapterwise** experiment. Results are averaged over novels (standard deviation). Best results are highlighted in **bold**.

| | CC | CQ |
|---|---|---|
| Semantics | 63.9 (15.8) | 54.4 (4.6) |
| STEL | 56.2 (15.6) | 52.7 (3.6) |
| Emotions | 53.4 (14.4) | 51.4 (3.1) |
| LUAR | **80.1** (10.0) | 53.5 (4.4) |

Table 3: AUC results of the **explicit** experiment. Results are averaged over novels (standard deviation). Best results are highlighted in **bold**. Results for the CQ evaluation are not highlighted because the large standard deviations prevent to chose a best model.

We also use AUC in this Character-Quote evaluation setup (CQ) to assess how well the target quotes spoken by character $c$ are ranked compared to quotes of other characters. Here, the AUC measures the probability that Equation 2 holds when randomly selecting a quote $q_i \in O_c$ and a quote $q_j \in \bar{O}_c$. Intuitively, a high AUC indicates that character representations are more similar to quote representations of the same character, thus showing that they contain useful information to attribute the right speaker to quotes.

## 4 Results

### 4.1 Chapterwise

Results for the chapterwise experiment are displayed in Table 2. In the CC evaluation setup, semantic-only representations built from the SBERT model appear to be quite good at distinguishing the voices of characters. We believe that SBERT particularly captures topical preferences, which appear as a useful discriminative feature of voices. Nonetheless, purely stylistic information seems to be worse at distinguishing voices than semantic-only embeddings, as suggested by STEL results. LUAR's high performance suggests that a combination of both content and stylistic information is desirable to achieve better and more stable discrimination among characters. Overall, the Emotions model seems quite misleading, as the AUC is the closest to random attribution (a random attribution would lead to an AUC of 50%).

When evaluating the capacity of these representations to attribute quotes, we see a drastically different picture. The performance of all models is just slightly higher than random attribution, indicating that the task is generally harder. This is not surprising, deciding which among thousands of quotes have been spoken by character $c$ given a corpus of around 10 quotes spoken by $c$ without access to contextual information is a challenging task, probably even for humans. Interestingly, the semantic-only baseline achieve the best results here. We hypothesise that the drop of performance of LUAR is mostly due to how it encodes quotes: it was trained to produce fine-grained author representations based on a corpus of multiple texts rather than to build rich text representations. In contrast, SBERT directly produces meaningful quote embeddings, leading to better performance for quotation attribution even if resulting character representations are less informative than LUAR's.

The high standard deviation in these results also suggests that distinguishing voices of characters is easier in some novels than in others. We analyze to which extent the semantic model and LUAR complement each other by looking at performance per novel in Appendix A and per character role in Appendix B.

### 4.2 Explicit

We present the results of the explicit experiment in Table 3. As expected, the performance of all models is worse than in the chapterwise experiment. Indeed, character representations built from explicit quotes have access to fewer quotes, reducing the amount of available information for each character. In the CC evaluation setup, LUAR still performs best at distinguishing voices of characters, followed by the SBERT model. Interestingly, even though queries are built with twice as few quotes on average than in the chapterwise experiment, we observe only a slight performance drop for LUAR and STEL. This observation suggests that explicit quotes constitute a strong signal of characters voice. However, we observe a larger drop for the SBERT model, indicating that there might be semantic variations between explicit quotes and other types of quotes. We hypothesize that such variations should occur less in stylistic cues of quotes, which is fur-

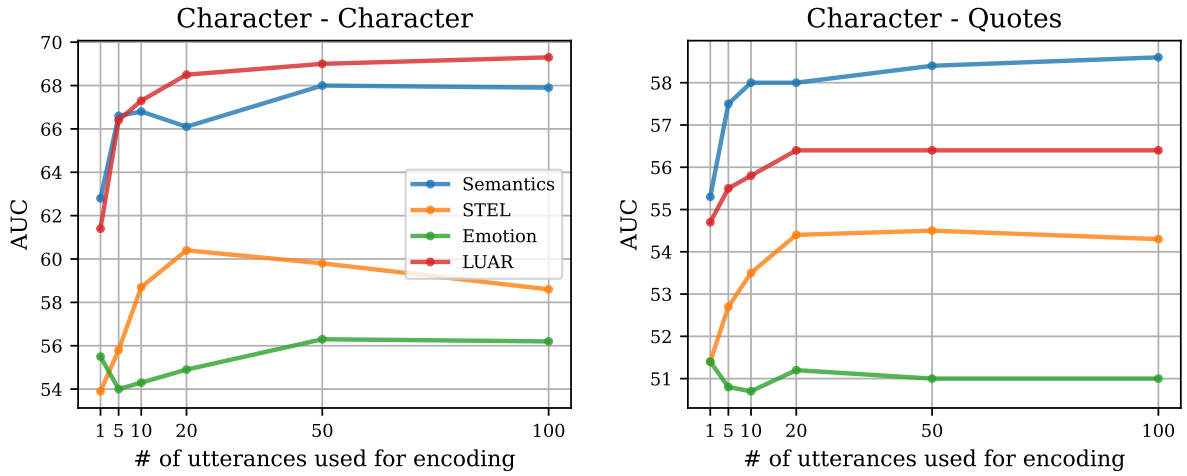


Figure 2: Results of the reading order experiment for the CC (left) and the CQ (right) evaluations. We look at AUC performance when varying the number of utterances used to create character representations.

ther supported by the lower AUC drop of STEL and LUAR.

Evaluating with the CQ setup, we can draw similar conclusions. Performances are worse than in the chapterwise experiment, and we see a larger AUC drop for the SBERT model than for models that encode style. However, the semantic model seems to remain the best at attributing unseen quotes on average, but it's not true for all novels.

Compared to the chapterwise experiment, we see very large standard deviations across novels. This is not surprising, some novels contain only a very small number of explicit quotes, leading to a smaller amount of queries as explained in Section 3.2. Although aggregated results suggest that the semantic model and the LUAR model still contain information for attributing quotes, when looking at results per novel, we observe that they can sometimes provide misleading attributions with AUC worse than 50%, but also provide a good ranking of quotes in other novels (the highest AUC is of 68% for *The Age Of Innocence* by Edit Wharton). Interestingly, we also observe a larger performance gap between the two models on some novels, indicating that they are more complementary when using explicit quotes only.

### 4.3 Reading Order

Results for the reading order experiment are displayed in Figure 2. Looking at the CC evaluation, we see that all models except Emotions have better performances when increasing the number of available utterances from 1 to 20. Increasing the number of quotes always improves the LUAR model, which successfully creates more fine-grained representations when accessing additional quotes. However, the STEL model peaks at 20 utterances, indicating that it can't really capture the style of characters with more quotes. Interestingly, the semantic model performance only varies slightly when using 5, 10 or 20 utterances, suggesting we can build meaningful semantic representations with a small number of quotes. This result further supports our hypothesis that the drop of performance between the Chapterwise and Explicit experiments is closely linked to semantic variations between explicit quotes and other types of quotes rather than simply due to lower query sizes. Overall, LUAR and Semantics build more informative representations using an increasing number of quotes from, 20 to 50 quotes of a character.

Results for the CQ evaluation show a similar trend, where the AUC of all models plateaus starting from 10 or 20 utterances, except for the Semantics that have increased performance with more data starting from 10 quotes. We hypothesize that stylistic information and topical preferences of characters can thus be captured by these models with a fairly low amount of quotes. A more complex understanding of characters does not always help to attribute quotes when using quote embeddings built with the same models, highlighting the need for additional contextual information.

## 5 Discussion

We conducted experiments to understand how character representations built from explicit quotes could help to improve quotation attribution. These quotes are particularly easy to attribute to characters and can thus be detected automatically to build *informative* character representations that can serve as additional inputs to quotation attribution systems. Results presented above suggest that explicit quotes might be a good proxy for the voice of fictional characters and that semantic and stylistic information of quotes can help attribute quotes. Nonetheless, we think that there might be semantic variations between explicit and other types of quotes and that adding stylistic information in representations of characters alleviates this shift.

Experiments conducted in this work are focused on *intermediate* and *major* characters, i.e. characters that participate more and shape the narrative. Although they represent a large number of different characters, *minor* characters often have less impact on the story and do not contribute significantly to the overall number of quotes that we want to attribute. However, even with characters that are more quoted, we observed discrepancies in authorial patterns of explicit quoting. While some authors quote all their characters explicitly at least 5 times in a chapter, some do not. As a result, we could build queries for only 53% of *intermediate* and *major* characters in the PDNC corpus. When looking at whole novels rather than at chapters, only 11% are explicitly quoted less than 5 times, among which 17% are major characters. Therefore, we can still build representations for a majority of characters, which motivated our work.

We studied stylistic information encoded in two off-the-shelf pretrained authorship verification models, LUAR and STEL. These models have been trained to distinguish thousands of authors of Reddit posts, and have been shown to transfer poorly to other domains (Rivera-Soto et al., 2021). Most novels do not contain stylistic traits captured by STEL, which probably explains why it is performing badly. We were aware of this limitation at first, but decided to test the model as an off-the-shelf solution to obtain stylometric representations. In the future, we plan to re-train a STEL-like model on literary texts such as drama. LUAR encodes both semantic and stylistic information, it is thus hard to infer the dimensions of content and style it captures as well as their respective contribution to the task. Its good performance on the Character-Character evaluation setup suggests that it gets dimensions of style that make sense in literature. More generally, interpretable authorship verification models (Patel et al., 2023) are an interesting direction as they combine the performance of neural approaches with the interpretability of frequency-based methods.

The high standard deviations across novels indicate that the task of distinguishing voices of characters is easier in some novels than in others. In the Chapterwise experiment (CC evaluation), the AUC of LUAR goes as low as 68% and as high as 91%. Ideally, we would like to understand the reasons behind these variations: Are some authors better at creating memorable voices? Is it easier in a particular genre? Interpretability and literary knowledge are key to answer these questions, that we leave for future work.

## 6 Conclusion

We presented a study of recent neural approaches to authorship verification applied to literary characters. We designed three experiments to assess if such models can be used to create meaningful character representations and to assess if explicit quotes were a good proxy of a character's voice. Our first evaluation focuses on the ability of these representations to distinguish characters, while our second quantifies the amount of information they contain to attribute unseen quotes. Results at the character level suggest that their voices are better distinguished when using a combination of stylistic and semantic information. Using style also helps to reduce the impact of the semantic shift observed between explicit quotes and other types of quotes. When attributing quotes, our results suggest that adding stylistic information does not necessarily improve over semantic-only models. We believe that the main cause is a poor domain transfer from Reddit to English novels. In the future, we plan to further analyze representations built from models trained on movie scripts (Azab et al., 2019; Li et al., 2023), which we argue should contain stylistic patterns more similar to the ones found in literary works. We also want to investigate how such representations can be incorporated into quotation attribution systems. Finally, we believe our approach could be used at a larger scale to investigate which authors/genre are better at constructing unique voices for their characters.

## References


Mahmoud Azab, Noriyuki Kojima, Jia Deng, and Rada Mihalcea. 2019. Representing movie characters in dialogues. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 99–109, Hong Kong, China. Association for Computational Linguistics.

Mikhail Bakhtin. 1984. *Problems of Dostoevsky's Poetics*. University of Minnesota Press.

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.

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.

Liviu P. Dinu and Ana Sabina Uban. 2017. Finding a character's voice: Stylome classification on literary characters. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 78–82, Vancouver, Canada. Association for Computational Linguistics.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1013–1019.

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

Hua He, Greg Kondrak, and Denilson Barbosa. 2010. The actor-topic model for extracting social networks in literary narrative.

Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. 2022. Learning and evaluating character representations in novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019, Dublin, Ireland. Association for Computational Linguistics.

Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023. Multi-level contrastive learning for script-based character understanding.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Grace Muzny, Mark Algee-Hewitt, and Dan Jurafsky. 2017a. Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32:ii31–ii52.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017b. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Ajay Patel, Delip Rao, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting llms.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 642–652, Online. Association for Computational Linguistics.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2019. Are fictional voices distinguishable? classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34, Minneapolis, USA. Association for Computational Linguistics.

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.

Krishnapriya Vishnubhotla, 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 (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Artjoms Šeļa, Ben Nagy, Joanna Byszuk, Laura Hernández-Lorenzo, Botond Szemes, and Maciej Eder. 2023. From stage to page: language independent bootstrap measures of distinctiveness in fictional speech.

## A Performance per Novel

We display the performance by novel for the Chapterwise experiment in Figure 3 and for the Explicit experiment in Figure 4. Note that for the Explicit experiment, novels 18 and 24 (*The Gambler* by Fyodor Dostoevsky (1887) and *The Sport of the Gods* by Paul Laurence Dunbar (1902) respectively) were not considered because we could not build queries due to the lack of explicit quotes. For the chapterwise experiment in the CQ evaluation setup, we see that LUAR's performances is higher than SBERT in 4 novels, indicating complementarity between these models. The picture is even more evident in the explicit experiment (CQ setup), where LUAR's outperformns SBERT in 8 novels. Overall, some novels exhibit characters voices where style information have more impact than on other novels.

## B Performance per Character Role

Table 4 displays results of the Chapterwise and Explicit experiments by character role. For the CC evaluation setup, LUAR performs very well on major characters, but struggles with intermediate characters. On the other hand, the semantic-only model performs better on intermediate characters. These results suggest complementarity between the two models, and that major characters exhibit more stylistic variations among them than intermediate characters. The latter result can be linked to the authorial process of creating memorable major characters, with more unique voices than intermediate characters.

For the CQ evaluation setup, it seems that all models are better at attributing quotes of intermediate characters, and we see a quite large gap between the two roles.

## C Computing information

We encode quotes with models on a 32-core Intel Xeon Gold 6244 CPU @ 3.60GHz CPU with 128GB RAM equipped with 3 RTX A5000 GPUs with 24GB RAM each. For each model tested, one GPU was enough to encode all quotes in the 28 novels. In total, running the full experiments took around 5 minutes for the Semantics and STEL models, 10 minutes for the Emotions model, and 1 hour for LUAR.

| | Chapterwise | | | | Explicit | | | |
|---|---|---|---|---|---|---|---|---|
| | **CC (M)** | **CC (I)** | **CQ (M)** | **CQ (I)** | **CC (M)** | **CC (I)** | **CQ (M)** | **CQ (I)** |
| Semantics | 62.9 (15.6) | **75.6** (12.8) | **53.1** (3.6) | **59.6** (5.0) | 58.0 (18.3) | **79.1** (16.9) | 51.8 (3.8) | **61.2** (7.3) |
| STEL | 55.4 (14.6) | 62.2 (11.1) | 52.2 (3.1) | 53.6 (3.3) | 52.5 (18.9) | 64.5 (23.7) | 51.5 (3.7) | 55.0 (10.5) |
| Emotions | 53.1 (15.1) | 59.5 (10.0) | 50.2 (3.2) | 53.6 (7.6) | 49.9 (18.2) | 59.6 (23.9) | 50.2 (3.2) | 53.6 (7.6) |
| LUAR | **91.2** (4.3) | 63.0 (12.7) | 52.1 (3.9) | 56.6 (4.9) | **87.6** (9.0) | 57.6 (25.1) | 51.6 (4.5) | 58.5 (7.3) |

Table 4: AUC results by character role for the Chapterwise and Explicit experiments. (M) means major and (I) intermediate.

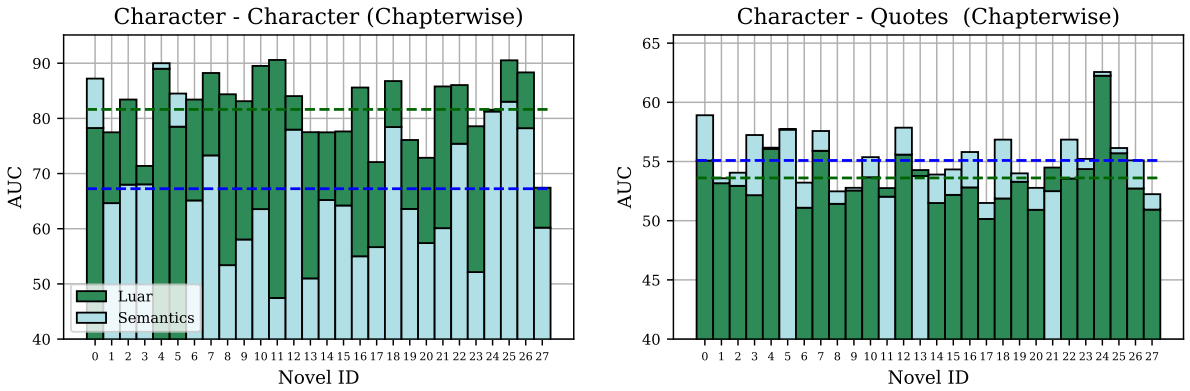


Figure 3: AUC per novel for the *Chapterwise* experiment.

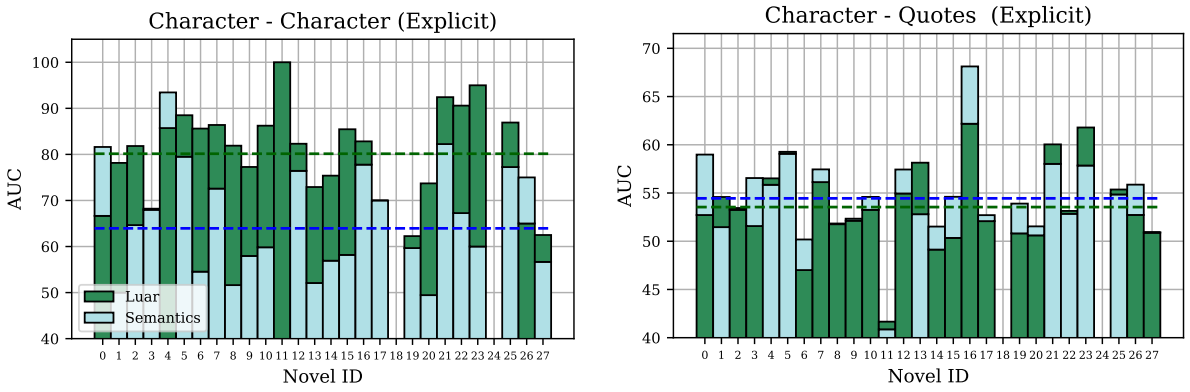


Figure 4: AUC per novel for the *Explicit* experiment.