Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5838–5848
Marseille, 20-25 June 2022

The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts

Krishnapriya Vishnubhotla[*†], Adam Hammond[‡], Graeme Hirst[*]
[*]Department of Computer Science, University of Toronto
[‡]Department of English, University of Toronto
[†]Vector Institute for Artificial Intelligence
vkpriya@cs.toronto.edu, adam.hammond@utoronto.ca, gh@cs.toronto.edu

Abstract

We present the Project Dialogism Novel Corpus, or PDNC, an annotated dataset of quotations for English literary texts. PDNC contains annotations for 35,978 quotations across 22 full-length novels, and is by an order of magnitude the largest corpus of its kind. Each quotation is annotated for the speaker, addressees, type of quotation, referring expression, and character mentions within the quotation text. The annotated attributes allow for a comprehensive evaluation of models of quotation attribution and coreference for literary texts.

Keywords: quotation attribution, literature, coreference

1. Introduction

Computational analysis of literary texts looks into Natural Language Processing (NLP) techniques to model aspects of narrative, events, and characters (Elsner, 2012; Bamman et al., 2014; Vishnubhotla et al., 2019). Past work in this area has focused mainly on analysing works of fiction, drawn from open-source platforms like Project Gutenberg[1] (Brooke et al., 2015; Bamman et al., 2020). The idiosyncrasies of literary text present several challenges to NLP models for named entity recognition, coreference resolution, character clustering, event detection, and speaker identification. The typical length of a text is several thousands of tokens, and the format and structure of the content vary widely depending on the genre, topic, time-period, and author of the text. Characters are referred to by various aliases, often incorporating notions of familial relations (*her father, Mr., Mrs., and Miss Bennet*) or social titles (*the baron*); mentions such as the former also can refer to different entities if used by different speakers (*my father*).

Consider, for example, the very first quotation in Jane Austen's *Pride and Prejudice*:

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Identifying the speaker of this quotation involves making several inferences: that the person being spoken to is Mr. Bennet, that the mention "his lady" refers to "Mr. Bennet's" lady, and that this is a proxy for Mr. Bennet's wife, who must be Mrs. Bennet (the first explicit mention of Mrs. Bennet is only in Chapter 2; several other characters are introduced to us in the meantime).

Previous work attempting to solve this task of identifying the speaker of a quotation in the text, or quotation attribution, has explored both rule-based systems (Glass and Bangay, 2007; Muzny et al., 2017) and machine learning models that are trained on an annotated dataset of quotations and speakers (Elson and McKeown, 2010; He et al., 2013; O'Keefe et al., 2012). Some of these approaches treat the task as a two-step problem, where quotations are first attached to mentions, and mentions are then attached to a canonical character name. The datasets developed for this task reflect this variation in methodology; some of them are annotated with quotation–mention–speaker information (Elson and McKeown, 2010; O'Keefe et al., 2012; Muzny et al., 2017), whereas others skip the intermediate mention annotation (He et al., 2013).

In this work, we present a new dataset, the Project Dialogism Novel Corpus (PDNC), comprising 22 full-length novels in which all quotations have been identified and annotated for speaker, addressees (who is being spoken to), characters mentioned, and the referring expression outside the quotation that indicates the speaker (if present). Our contributions are as follows:

- PDNC is by an order of magnitude the largest dataset of annotated quotations for literary texts in English, in terms of the number of tokens covered, the number of annotated quotations and characters, as well as the number of character mentions (even though we limit ourselves to mentions within quotations).
- We release, along with the dataset, a comprehensive set of annotation guidelines that cover several idiosyncrasies of literary texts, and which we hope will help standardize future annotation work in this domain.
- We evaluate two state-of-the-art quotation attribution systems on this dataset, which obtain average accuracies of 0.62 and 0.63 respectively. We also evaluate a simple semi-supervised classification baseline that achieves competitive results.

[1]https://www.gutenberg.org/

- We use our annotations to analyze the performance of these models and pinpoint common failure points, which will help inform future work in this area.

All data and code associated with this work will be made publicly available at `https://github.com/Priya22/pdnc-lrec2022`.

2. Background

We review past datasets of quotations in the literary domain, as well as automatic models for the task of quotation attribution.

2.1. Prior Datasets

The Columbia Quoted Speech Attribution (CQSA) corpus from Elson and McKeown (2010) contains annotations for 3176 instances of quoted speech from 4 novels by each of 4 authors, and 7 short stories from 2 others; only parts of the full-length novels are annotated. Quotations are annotated at the mention-level, i.e, the speaker is chosen from a set of candidate mentions that occur in the nearby context. These mentions are then resolved to speakers by using an off-the-shelf coreference tool. He et al. (2013) annotate a dataset of three novels, *Pride and Prejudice*, *Emma*, and *The Steppe*; the latter two are also present in the CSQA corpus. Their annotation method links quotations directly to canonical characters, rather than mentions. Muzny et al. (2017) released the QuoteLi dataset, comprising 3103 quotations annotated with both mention and speaker information. The quotations are drawn from the same three novels as those of He et al. (2013). Finally, Sims and Bamman (2020) annotate the first 2000 tokens of 100 novels from the LitBank dataset[2]. Quotations are linked to a unique speaker from a predefined list of entities. Though this dataset spans the largest number of novels (100), the restricted range of tokens considered results in only 1765 total annotations.

LitBank also contains annotations for coreference, for the same set of 2000 tokens across 100 novels. A total of 29,103 tokens are annotated, of which 24,180 refer to a person, and the rest to other named entities such as places, organizations, vehicles, etc (Bamman et al., 2020). Prior to this, Vala et al. (2016) annotated coreference in *Pride and Prejudice*.

2.2. Models of Quotation Attribution

Elson and McKeown (2010) proposed a classification approach for quotation attribution that classifies quotations into one of several types based on whether the speaker is explicitly indicated by an adjoining expression (explicit), appears without an attribution (implicit), is indicated by an anaphoric mention, is part of a dialogue chain, etc (see Table 1 for examples of each quotation type). A separate classifier is trained for each of these cases, taking as input a feature vector that encodes information relating to positions of mentions and quotations surrounding the target. Their model achieves an accuracy of 83% on their dataset, but uses gold labels as part of the pipeline.

O'Keefe et al. (2012) treat the task as a sequence decoding problem, where the set of speaker attributions in a document is treated as a text sequence to be predicted; i.e, the decision for the current quotation is made based on the previous n attribution labels. While this method works well for news data, it fails to beat a rule-based baseline for literary texts. He et al. (2013) approach quotation attribution as a ranking problem between candidate speakers; their SVM-based ranking model selects a speaker based on a feature vector comprising contextual and topic information.

Muzny et al. (2017) describe a two-step process for quotation attribution, where quotations are first linked to mentions, and mentions to entities. Each step is composed of a set of deterministic sieves, designed to capture cases of increasing complexity. For example, the first sieve looks for explicit trigram patterns of Quote–Speech Verb–Mention. This system is described further in Section 5.1.1.

Hammond et al. (2020) describe a semi-supervised classification approach to quotation attribution that, similar to those of Elson and McKeown (2010) and He et al. (2013), builds a feature vector and trains a classifier to predict the speaker. Their features are based primarily on lexical and syntactic features drawn from work in computational stylometry, and uses an iterative classification approach where high-confidence predictions of the classifier are repeatedly incorporated into the training set.

3. The Project Dialogism Novel Corpus

We draw our novels from open-source texts available on the Project Gutenberg platform. In selecting these novels, our aim has been to annotate texts in a variety of genres (literary fiction, children's literature, detective fiction, and science fiction are represented); from the LitBank and QuoteLi corpora, to facilitate comparison and validation; and of broad interest to a variety of scholars while still relevant to our group's interest in stylistic diversity and dialogism (Hammond et al., 2020; Vishnubhotla et al., 2019). Further, we have chosen to annotate multiple novels by Jane Austen, in order to facilitate comparative analysis of a single author's oeuvre (Austen was chosen because she is included in all existing corpora).

3.1. Annotated Attributes

Each quotation in our corpus of texts is annotated with the following attributes:

1. **Speaker:** The character uttering the quotation. We limit each quotation to having a single speaker; certain special cases are highlighted in Section 4.4.

[2] https://github.com/dbamman/litbank

Quotation	Annotations
"You must not be too severe upon yourself," replied **Elizabeth**	**Speaker:** Elizabeth Bennet **Addressees:** (Mr. Bennet, Kitty) **Quote type:** Explicit **Referring Expression:** replied Elizabeth **Mentions:** ('you', Mr. Bennet), ('yourself', Mr. Bennet)
With an air of indifference **he** soon afterwards added: *"How long did you say he was at Rosings?"*	**Speaker:** George Wickham **Addressees:** Elizabeth Bennet **Quote type:** Anaphoric **Referring Expression:** he soon afterwards added **Mentions:** ('you', Elizabeth Bennet), ('he', Colonel Fitzwilliam)
"But not before they went to Brighton?"	**Speaker:** Elizabeth Bennet **Addressees:** Jane Bennet **Quote type:** Implicit **Referring Expression:** **Mentions:** ('they', [George Wickham, Lydia])

Table 1: Annotations for three sample quotations from PDNC, one for each quotation type. The speaker in each example is highlighted in bold, and mentions within quotations are underlined.

2. **Addressee(s):** The set of character(s) being addressed by the speaker. This includes any character that is in the vicinity of the speaker and can "hear" the uttered quotation.

3. **Quotation Type:** Following previous work, we distinguish between explicit, anaphoric, and implicit quotations. See Table 1 for an example of each.

4. **Referring Expressions:** For explicit and anaphoric quotations, we obtain the part of the text that indicates who the speaker is, the verb for the action of speaking, and sometimes, also the addressees.

5. **Mentions:** Finally, we also annotate all characters who are mentioned within a quotation, either explicitly by name or through a pronoun or pronominal phrase. Each mention is linked to the character or set of characters that it refers to.

In addition, each novel is also annotated with a list of characters present in the novel. Each character is associated with a "main name" (e.g., Elizabeth Bennet), as well as a set of aliases by which they are referred to in the text (e.g., Lizzy, Liz, Elizabeth). The character list includes any character who either speaks, is addressed, or is mentioned in a quotation; therefore we also have characters who are never explicitly assigned a proper name, such as "The Old Man in the Crowd".

3.2. Dataset Statistics

We list key characteristics of PDNC in Table 2. A total of 35,978 quotations are identified and annotated for the attributes listed in Section 3.1. On average, we have 1.79 aliases per character, and 1.82 mentions annotated per quotation. Of the 992 characters in our character lists, 655 are speakers of a quotation; of these, 321 characters can be classified as "minor", having 10 or fewer spoken quotations. Margaret Schlegel from *Howards End* is the most loquacious character across all novels, with 1040 quotations, followed by Jake Barnes from *The Sun Also Rises*, Katherine Hilbery from *Night and Day*, and Anne Shirley from *Anne of Green Gables*.

Figure 1 in Appendix A shows the distribution of quotation types across all novels. We see that implicit quotations make up the largest percentage of annotations (∼37%), followed by explicit (∼33%) and anaphoric (∼29%) quotation types, though the distribution shows a large spread. *Alice in Wonderland* consists mostly of explicit quotations (84%), whereas Dostoevsky's *The Gambler* is at only 12%.

We note that PDNC is by far the largest dataset of annotated quotations for works of English Literature. A comparison with previous datasets is presented in Table 3. Even though we annotate only for mentions within quotations, our count of 62,587 mention annotations is much larger than LitBank's 29,103.

PDNC also contains the largest number of tokens per document (79,745), since we annotate entire novels rather than portions of each. We think that this is an invaluable resource for several open problems in the computational analysis of literature, allowing for tracking character mentions across larger spans of text, studying changes in character style, emotions, and character networks throughout the course of a novel, and the variation of each of these with author and genre.

4. PDNC: The Annotation

In this section, we describe our annotation process, from developing the guidelines to preprocessing the texts, the annotation platform, and how we resolved disagreements between annotators.

Novel	Author	# Tokens	# Quotations	# Characters	# Mentions
A Handful Of Dust	Evelyn Waugh	70299	2617	104	3198
A Room With A View	E. M. Forster	67434	1989	67	3111
Alice's Adventures in Wonderland	Lewis Carroll	26826	1048	51	683
Anne Of Green Gables	Lucy Maud Montgomery	103291	1779	114	5168
Daisy Miller	Henry James	22007	725	10	1021
Emma	Jane Austen	161070	2116	18	6318
Howards End	E. M. Forster	112674	3131	56	4358
Night And Day	Virginia Woolf	170706	2800	54	3575
Northanger Abbey	Jane Austen	78081	1017	20	2358
Persuasion	Jane Austen	83695	702	35	2186
Pride And Prejudice	Jane Austen	122692	1708	77	4797
Sense And Sensibility	Jane Austen	120810	1545	25	4676
The Age Of Innocence	Edith Wharton	103062	1600	55	2556
The Awakening	Kate Chopin	50234	738	22	981
The Gambler	Fyodor Dostoevsky (Trans. C.J. Hogarth)	61508	1068	26	2057
The Invisible Man	H. G. Wells	49956	1274	33	926
The Man Who Was Thursday	G. K. Chesterton	58352	1357	31	1700
The Mysterious Affair At Styles	Agatha Christie	57302	2226	30	3485
The Picture Of Dorian Gray	Oscar Wilde	80483	1501	45	3336
The Sign of the Four	Sir Arthur Conan Doyle	43872	891	36	1784
The Sport Of The Gods	Paul Laurence Dunbar	41470	830	38	1524
The Sun Also Rises	Ernest Hemingway	68585	3316	45	2789
Total		1754409	35978	992	62587

Table 2: The set of novels annotated in PDNC, with the number of annotated quotations, characters, and mentions in each.

Corpus	CQSA (2010)	He et al. (2013)	Muzny et al. (2017)	LitBank (2020)	PDNC (2021)
# Texts	6	3	3	100	22
# Quotations	3176	1901	3103	1765	35978

Table 3: A comparison of PDNC with previous datasets for quotation attribution in literary texts.

4.1. Annotation Platform

We designed our annotation platform from scratch as a web-based interface. A screenshot of the interface is shown in the Appendix, Figure 2. The main components include the character list, which allows the annotator to add and remove characters and associated aliases; the text box, which highlights quotations and mentions within the text (different color codes indicate the type and annotation status of the quotation or mention spans); and the annotation area, where values for the desired attributes of a quotation or mention can be set by the annotator. The platform also includes an interface that takes as input two sets of annotations of the same text and generates a file with any disagreements that occur for an annotated attribute, including mis-matches in character lists.

4.2. Annotation Process

All our annotators were university-level literature students familiar to one of the authors. Each novel in our corpus was annotated separately by two annotators, and the resulting annotations were then compared to generate a list of "disagreements". Disagreements were grouped by quotation, and occur when the annotations do not match for any of the attributes listed in Section 3.1. The two annotators then went through a consensus exercise, where they discussed all disagreements, re-annotated the relevant quotations, and once again checked for disagreements (in practice, no more than three rounds of consensus were necessary).

4.3. Pre-processing the texts

The raw text for each novel is obtained from the Project Gutenberg platform. This is then processed using the GutenTag software[3] from Brooke et al. (2015), which outputs an initial list of characters and aliases, and also identifies quotations within the text. We also pre-identify mentions within each quotation by looking for occurrences of any character names, aliases, or words from a predefined list of pronouns.

[3]https://gutentag.sdsu.edu/

4.4. Annotation Guidelines

The complexity of narrative structure and style of literary novels means that several ambiguities can arise while determining any of the annotated attributes. We developed a comprehensive set of guidelines that attempt to cover as many as possible of the cases that we came upon in our texts. These guidelines underwent several revisions as we progressed through different novels, and were informed by feedback from our annotators as well as the authors of this work. We make the complete set of guidelines publicly available and hope it will help guide future work in this area. We highlight a few interesting cases below:

- Special aliases: Narrators of first-person narratives receive the special alias "_narr"; when more than one character speaks a quotation in unison, it is attributed to "_group"; when the identity of the speaker is unknowable in context, it is attributed to "_unknowable".
- Multiple addressees: In situations in which many characters are present, our guidelines designate an addressee as anyone "whom the speaker seems to believe can hear them."
- Locating referring expressions: Our guidelines include explicit instructions for annotating referring expressions in cases in which they are difficult to annotate, in which they introduce long or multi-part quotations, and in which multiple referring expressions are applied to single quotation.

5. Quotation Attribution

We now turn our focus to the analysis of quotation attribution models, where the task is to identify "who said what". Building an automated attribution system from scratch is generally a multi-step process: we first need to identify quotations in the text, build a list of characters and their aliases, and then attribute each quotation either directly to character, or first to a mention followed by an additional coreference resolution step to identify the associated character.

5.1. Review of Existing Systems

We briefly describe two models for quotation attribution that are the current state-of-the-art.

5.1.1. A Two-Stage Sieve Approach

Muzny et al. (2017) propose a deterministic, two-step, approach to quotation attribution that relies on several sieves of increasing complexity to first link each quotation to a mention, and then link the mention to a character entity. The latter step involves applying a coreference resolution model to the text. Since our focus is primarily on the quotation attribution, we briefly describe the main sieves associated with the first step:

1. **Trigram Matching (Tri-1)**: This identifies patterns of the type Quote-Mention-Speech Verb, or Quote-Speech Verb-Mention, to extract quotations where the speaker is indicated by the associated referring expression (e.g., *"she said"*, or *"said Elizabeth"*).
2. **Dependency Parses (Dep-2)**: This inspects dependency parses of sentences on either side of the target quotation for speech verbs with an `nsubj` relation that points to a character mention.
3. **Single Mention Detection (Single-3)**: This looks for instances where there is only a single mention in the non-quotation text of the associated paragraph, and attributes the quotation to that mention.
4. **Vocative Detection (Voc-4)**: This looks for vocative patterns involving mentions in the previous quotation (e.g., *"are you sure, Lizzy?"*), and links the quotation to the the associated mention.
5. **Paragraph Final Mention (Par-5)**: This attributes a quotation occurring at the end of a paragraph to the final mention of the previous sentence.
6. **Conversational Pattern (Conv-6)**: This looks for consecutive sequences of quotations (i.e, uninterrupted by non-quote text), and links an unattributed quotation to the speaker of the quotation two steps behind. Muzny et al. specify a less-restricted version of this where the requirement of "uninterrupted by non-quote text" is removed.

The sieves, in order, deal with quotations in increasing order of the difficulty of attribution. The easy cases, such as explicit and most anaphoric quotations, are captured by the first two sieves; the latter ones deal with the more complex, implicit quotations that require additional knowledge of the surrounding context.

5.1.2. BookNLP

BookNLP[4] is a tool for natural language processing of literary texts (and other long documents) in English. The pipeline performs, among other things, dependency parsing, named entity recognition, coreference resolution, quotation attribution, and referential gender inference. The latest version of BookNLP is trained on LitBank's annotations of character entities (Bamman et al., 2020) and quotations (Sims and Bamman, 2020). While the exact model for quotation attribution is not described in a publication, we infer from the code that it uses a BERT-based model that takes as input the quotation text and its surrounding context, and links each quotation to a character mention. Mention-to-entity resolution is performed by a separate pipeline step that precedes quotation attribution.

Muzny et al. (2017) specify that they use BookNLP's coreference resolution system for the mention–entity step of their pipeline, though at the time of publication,

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

the latest version of BookNLP was not yet released. Since our focus here is on evaluating models of quotation attribution, separately from coreference resolution, we plug in the latest outputs of BookNLP's coreference resolution system into the two-stage attribution approach of Muzny et al. (2017).

5.1.3. A Semi-Supervised Stylometric Approach

One of the key uses of our corpus is in work on dialogism, i.e. variation in the speaking styles of characters in a novel as compared to one another and to the narrator. Hammond et al. (2020) propose a stylometric, semi-supervised classification approach to quotation attribution that relies on the stylistic characteristics of the quotation text to identify the speaker. We test here a slightly modified version of that approach, the details of which are in Appendix B.

Briefly, for each quotation in our dataset, we extract a set of features based on the annotated attributes: the text of the quotation, the referring expression (if present), and the set of mentions. These features are drawn from prior work in computational stylometry (Altakrori et al., 2021; Vishnubhotla et al., 2019). The feature vectors are passed to a classifier that is trained to predict the the speaker in an n-way classification setup. The model follows a semi-supervised approach that iteratively extracts high-confidence predictions from the test set and adds them to the training set for the next round of classification.

Note that this model does not function as a stand-alone quotation attribution system, since it assumes access to both a fixed list of characters and gold speaker labels. We merely test it on the PDNC dataset to examine the viability of a stylometric approach to the speaker attribution problem, and as a complement to existing approaches.

5.2. Experimental Setup

To test the Muzny et al. (2017) model, we use the Python re-implementation from (Sims et al., 2019), as it fits well with our Python pipeline (the original implementation is in Java, and integrated with the StanfordCoreNLP pipeline). We use the latest version of BookNLP to identify quotations within the text and a list of character clusters. The latter is obtained as an output of the coreference resolution module, which clusters together mentions within the text that are presumed to refer to the same entity. As such, the model does not build a list of canonical character names with which to associate quotations; rather, each quotation is attributed to an entity cluster.

This presents a slight problem for evaluating their performance based on our gold standard annotations. Consider a character cluster identified by BookNLP as follows: $\{$*my, her, mingott, i*$\}$. This is identified as the speaker of a quotation, whose speaker label in PDNC is *Mrs. Lovell Mingott*. However, we also have another character in the same novel, *Uncle Lovell Mingott*. The ambiguity in matching characters can in this particular case be resolved by inferring the gender of both characters via the associated pronouns; however, we observed quite a few cases where either a resolution was not possible (e.g., the name *Mingott* can refer to any member of the Mingott family; even *Miss Mingott* could refer to more than one unmarried female of the Mingott family, depending on the context), or the character cluster contained conflicting pronouns (both *her* and *him* appeared along with the name *Mingott*).

In our evaluation of the systems, we do not consider the cases where this ambiguity could not be resolved; this results in the evaluation size being smaller than the set of identified quotations. We report the difference in these sizes in Table 4; for most novels, this number lies in the lower hundreds.

For the stylometric classification model, we limit ourselves to characters with at least 10 annotated quotations, in order to avoid the long tail of minor characters. To further mitigate the class imbalance issue, we oversample from the minority classes. We use a Logistic Regression classifier with a grid search over the regularization hyperparameter. The initial training and test sets for each novel are based on quotation types: explicit quotations are assigned to the training set, and the rest form the test set. The probability threshold for each iteration of the classification is dynamically determined as a mean of the probabilities over correct predictions; we observed that this parameter varies from novel to novel, and that the classification setup is quite sensitive to this value.

5.3. Results

Table 4 shows the performance of the two state-of-the-art (SoTA) attribution systems and the stylometric classifier on the PDNC novels. Note that, for the former, since we use BookNLP as a common pipeline for the quotation identification and character name clustering steps, both systems are evaluated on the same set of quotations. For the stylometric model, the accuracy is calculated on the set of non-explicit quotations by non-minor characters (at least 10 annotated quotations).

We see that there is a large variation in the performance across novels, for all models. Certain novels, such as *Anne of Green Gables*, seem easier to attribute for both of our SoTA models; likewise, others seem to present difficulties across the board (*The Sport of the Gods*). *Alice in Wonderland*, in particular, achieves near-perfect accuracy scores. This can partly be attributed to the fact that Alice is by far the most common speaker, contributing to nearly 42% of all quotations in the novel, and nearly 84% of the quotations are explicit.

With the stylometric model, which is evaluated only on non-explicit quotations, we do quite well on certain novels that the SoTA systems struggle with. For *The Age of Innocence*, for example, the stylometric model correctly attributes 75% of the implicit and anaphoric quotations, which account for nearly 80% of the total annotated quotations. By contrast, both the Muzny and

	State-of-the-art				Stylometric	
Novel	**# Identified**	**# Eval**	**Muzny et al.**	**BookNLP**	**# Eval**	**Stylo**
A Room With A View	2071	1857	0.58	**0.59**	1424	0.57
Alice In Wonderland	1122	965	**0.95**	0.93	157	0.76
Anne Of Green Gables	1841	1726	**0.88**	0.86	660	0.60
Daisy Miller	749	713	0.70	**0.74**	390	0.73
Emma	2108	1935	0.61	**0.62**	1422	0.60
Handful Of Dust	2732	2502	0.58	**0.59**	1956	0.54
Howards End	3304	2917	0.61	**0.66**	2195	0.56
Night And Day	2901	2619	**0.74**	0.72	1776	0.68
Northanger Abbey	1072	1001	**0.59**	0.54	734	0.66
Persuasion	786	655	**0.69**	0.63	330	0.33
Pride And Prejudice	1779	1681	0.63	**0.64**	1133	0.48
Sense And Sensibility	1546	1472	0.63	**0.64**	886	0.31
The Age Of Innocence	1912	1466	0.44	**0.45**	1235	0.75
The Awakening	782	705	0.59	**0.62**	517	0.65
The Gambler	1128	1012	0.40	**0.42**	920	0.74
The Invisible Man	1277	1103	**0.80**	0.79	585	0.52
The Man Who Was Thursday	1339	1264	**0.78**	0.76	479	0.44
The Mysterious Affair At Styles	2228	2103	**0.50**	0.42	1791	0.66
The Picture Of Dorian Gray	1539	1450	0.59	**0.66**	1068	0.63
The Sign Of the Four	900	815	0.42	**0.44**	710	0.72
The Sport Of The Gods	885	783	0.46	**0.50**	625	0.44
The Sun Also Rises	3324	3219	0.52	**0.55**	2223	0.65
Total	37325	33963	0.62	**0.63**	23216	0.59

Table 4: Accuracy scores for the quotation attribution systems from Muzny et al., BookNLP, and the stylometric classifier (Stylo). The first numerical column in each row for the SoTA models is the number of quotations identified by BookNLP, the second is the number of quotations for which the predicted cluster of speaker mentions could be matched with our annotated list of characters. For the Stylo model, # Eval is the number of non-explicit quotations by major characters in PDNC for that novel.

	Explicit		Anaphoric		Implicit	
Method	**# Qs**	**Acc.**	**# Qs**	**Acc.**	**# Qs**	**Acc.**
Muzny et al.	11545	0.96	9855	0.48	12551	0.41
BookNLP	11545	0.94	9855	0.46	12551	0.46
Stylo	11556	–	10072	0.67	13133	0.54

Table 5: Breakdown of the performance of our models by quotation type. Stylo refers to the stylometric model.

BookNLP models achieve accuracies of about 45%.

5.3.1. Performance by Quotation Type

Table 5 presents a breakdown of the performance of each of our three models by quotation type. Note that since we use explicit quotations as the training set for the stylometric system, we do not report an accuracy score in that cell. Both the Muzny and BookNLP models perform quite well on explicit quotations. As expected, implicit quotations are the hardest to attribute. That anaphoric quotations do not fare much better indicates that the coreference resolution part of the attribution pipeline is responsible for many mis-attributions; we verify this hypothesis in the next section.

5.4. Evaluating the Sieves

We examine how often the heuristic sieves proposed by Muzny et al. (2017) hold up across all our novels. For each sieve, we try to answer questions with regard to the number of quotations of each type captured by the sieve, it's performance on these quotations, and the possible reasons for mis-attributions.

We first take a qualitative look by examining the performance of the model for one of the novels in our corpus, *The Age of Innocence*. Table 6 lists examples quotations from the text that are wrongly attributed based on mentions in the context surrounding the target quotation; many of these occur due to sentence structures that are not straightforward. We observe several such instances in this text and others, indicating that the surrounding contextual information alone may not always be sufficient to attribute quotations.

However, the most common source of attribution errors that we observed in our analyses was failure of the coreference resolution module. Even with BookNLP's state-of-the-art model, a large number of character clusters either are not associated with a character en-

Sieve	Example
Tri-1	"Gad," Archer heard Lawrence Lefferts say, "*not one of the lot holds the bow as she does*" and Beaufort retorted "Yes but that's the only kind of target she'll ever hit." **Speaker:** Lawrence Lefferts **Predicted:** Julius Beaufort
Dep-2	Mr. Welland, beaming across a breakfast table miraculously supplied with the most varied delicacies, was presently saying to Archer *"You see, my dear fellow, we camp we literally camp. I tell my wife and May that I want to teach them how to rough it."* **Speaker:** Mr Welland **Predicted:** Newland Archer
Voc-4	Q1 (Newland): "Your mother?" Q2: *"Yes the day before she died."* **Speaker:** Dallas Archer **Predicted:** your mother

Table 6: Example quotations from *The Age of Innocence* that are mis-attributed by sieves of Muzny et al.'s attribution model. The quotation under consideration is italicised.

	Explicit		Anaphoric		Implicit	
Sieve	**# Qs**	**Acc.**	**# Qs**	**Acc.**	**# Qs**	**Acc.**
Tri-1	6500	0.98	5750	0.50	24	0.42
Dep-2	3529	0.96	2872	0.46	271	0.43
Single-2	1212	0.82	831	0.26	993	0.43
Voc-4	55	0.29	68	0.22	1669	0.46
ParFinal-5	0	–	1	1.00	6	0.33
ConvPat-6	156	0.24	207	0.27	7208	0.52
BASE-7	93	0.26	126	0.29	2380	0.30

Table 7: Breakdown of the performance of each sieve from Muzny et al. (2017) by quote type. **# Qs** indicates number of quotations.

tity explicitly by name (e.g., {*you, yourself, your, i, she, her*} forms one of the clusters), or mix together mentions of several different characters into a single cluster, sometimes with opposing gendered pronouns ({*herself, my, yourself, archer, his*}).

5.4.1. Quantitative Analysis

Table 7 details the performance of each of the 6 sieves, along with an additional baseline sieve, BASE-7 (attribute to the most common mention in a 5000-word window surrounding the target quotation), when divided by quotation type. Surprisingly, not all explicit quotations are captured by the trigram and dependency parse sieves, indicating the prevalence of more-complex referring expressions, even with explicit character mentions. We also note that the Paragraph Final Mention sieve rarely comes into play. The accuracy on anaphoric and implicit quotations doesn't exceed 50% across the board, highlighting again the key role played by the coreference resolution module.

5.5. Discussion

Our results demonstrate the challenges posed by literary novels for quotation attribution. Accuracy scores vary widely across novels for all three models that we evaluate. Implicit and explicit quotations in particular are hampered by the mention-to-entity step of the pipeline, due to the much harder task of coreference resolution in this domain. The stylometric model, which directly predicts speaker labels, does relatively better on these subsets. Though most recent work in this area has moved away from building canonical character lists, instead defaulting to mention clusters, we think that the former approach is better for a standardized evaluation of the task. It is also beneficial for downstream applications that use these outputs, such as analyzing stylistic patterns of individual characters, building networks of speaker interactions, and analyzing broader trends in these across authors and genres.

6. Conclusion

We presented a new dataset of quotation annotations for English literary texts, with 35,978 quotations across 22 full-length novels annotated for speaker, addressees, quotation type, referring expression, and mentions. This is the largest dataset of quotations and mentions in this domain. We hope that the comprehensive set of annotation guidelines developed as part of the annotation process will be useful for any future work in this area. We hope to expand PDNC with a more diverse set of texts in the future.

We demonstrated that existing quotation attribution models still have a long way to go in reliably identifying the speaker of a quotation, despite being trained on literary datasets. PDNC provides a new source of training data for these models, and its annotated attributes are also useful in identifying the causes of errors in attribution. We showed that a stylometric classification model serves as a competitive baseline for the task, and would be a useful complement to attribution models.

Acknowledgements

We are grateful to the annotators involved in the creation of this dataset, Alanna Carolan, Sofia Chabchoub, Leah Duarte, Bisman Kaur, Sol Kim, Sanghoon Oh, Jovana Pajovic, and Beck Siegal. The first author is supported by funding from the Natural Sciences and Engineering Research Council of Canada, and resources provided by the Vector Institute of Artificial Intelligence. This research was supported by a Connaught New Researcher Award from the University of Toronto.

7. Bibliographical References

Altakrori, M., Cheung, J. C. K., and Fung, B. C. (2021). The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256.

Bamman, D., Underwood, T., and Smith, N. A. (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.

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

Brooke, J. and Hirst, G. (2013). A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–679.

Brooke, J., Hammond, A., and Hirst, G. (2015). GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47.

Elsner, M. (2012). Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644.

Elson, D. K. and McKeown, K. R. (2010). Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Glass, K. and Bangay, S. (2007). A naive salience-based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA'07)*, pages 1–6.

Hammond, A., Vishnubhotla, K., and Hirst, G. (2020). The words themselves: A content-based approach to quote attribution. In *Proceedings of the Digital Humanities 2020 Conference*.

He, H., Barbosa, D., and Kondrak, G. (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.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Muzny, G., Fang, M., Chang, A., and Jurafsky, D. (2017). 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.

O'Shea, J. (2013). Alphabetical order 277 word new function word list. `https://semanticsimilarity.files.wordpress.com/2013/08/jim-oshea-fwlist-277.pdf`. [Retrieved Oct. 2019].

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Sims, M. and Bamman, D. (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.

Sims, M., Park, J. H., and Bamman, D. (2019). Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Vala, H., Dimitrov, S., Jurgens, D., Piper, A., and Ruths, D. (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.

Vishnubhotla, K., Hammond, A., and Hirst, G. (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.

Appendix

A. Types of Quotations in PDNC

Figure 1 shows box-and-whisker plot of the three quotation types annotated in our dataset. The central region (the box) indicates the "middle portion" of the data distribution, i.e, the range covered between the first quartile (the 25% mark) and the third quartile (the 75% mark), with the median (50% mark) lying at line inside the box. The whiskers, the dashes on either end of the plot, are at a distance of 1.5 times the inter-quartile length (inter-quartile length is the distance between the first and third quartiles). Points beyond the whiskers are considered outliers.

B. Stylometric Classification for Quotation Attribution

Here, we describe the classification-based approach to quotation attribution adapted from Hammond et al. (2020).

Lexical Features — Character-Level	**Lexical Features — Word-Level**
1. Characters count (N)	1. Tokens count (T)
2. Ratio of digits to N	2. Average sentence length (in characters)
3. Ratio of letters to N	3. Average word length (in characters)
4. Ratio of uppercase letters to N	4. Ratio of alphabets to N
5. Ratio of tabs to N	5. Ratio of short words to T (a short word has a length of 3 characters or less)
6. Frequency of each alphabet (A-Z), ignoring case (26 features)	6. Ratio of words length to T. Example: 20% of the words are 7 characters long. (20 features)
7. Frequency of special characters: <>%|{} []/\@#˜ +-*=$ˆ &_()’ (24 features).	7. Ratio of word types (the vocabulary set) to T
Syntactic Features	
1. Frequency of Punctuation: , . ? ! : ; ’ ” (8 features)	
2. Frequency of function words from O’Shea (2013) (277 features)	

Table 8: List of stylometric features from Altakrori et al. (2021)

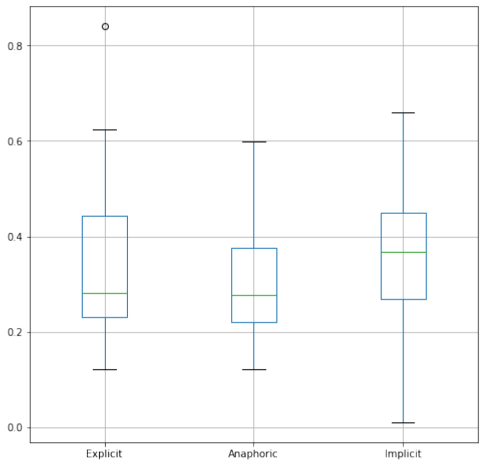

Figure 1: Distribution of quotation types across novels in PDNC.

Feature Extraction

The feature vector is composed of the following sets of features:

1. **Stylometric features:** From the quotation text, we extract a set of 371 features that capture character and word-level lexical and syntactic features of the text. These features were drawn from prior work in authorship attribution and computational stylometry, particularly that of Altakrori et al. (2021). The list of features is in Table 8.

2. **TF-IDF Counts:** We vectorize the quotation text (excluding stop words used in Feature Set 1 above), the text of the referring expression, and the set of entities mentioned within the quotations, where available, using TF-IDF counts.

3. **Lexicon-based features:** For each quotation text, we find the average value of the words in the quotation along a set of lexical dimensions, where the values are obtained via lexicons. The first set of features are the six dimensions of style as described by Brooke and Hirst (2013) — literary, abstract, objective, colloquial, concrete, subjective, polarity — with the associated lexicon provided by the authors. The second set of features comes from the NRC Emotion Intensity Lexicon, which associates each word with a real-valued score along eight basic emotions — anger, anticipation, fear, joy, sadness, and trust — and two sentiments, positive and negative (Mohammad and Turney, 2013). Finally, we compute these features for the three emotion dimensions of valence, arousal, and dominance, the lexicons for which are obtained from the work of Mohammad (2018).

Classification

Our classification model is a semi-supervised approach that iteratively extracts high-confidence predictions from the test set and adds them to the training set for the next round of classification. Let us assume a dataset of quotation–speaker pairs (X, y), and an initial train and test set of quotation–speaker pairs $(X_{train\_init}, y_{train\_init})$ and $(X_{test\_init}, y_{test\_init})$. The classification pipeline proceeds as follows:

1. Set $(X_{train}, y_{train}) \leftarrow (X_{train\_init}, y_{train\_init})$ and $(X_{test}, y_{test}) \leftarrow (X_{test\_init}, y_{test\_init})$.

2. Extract feature vectors for X_{train} and X_{test}.

3. Train a classifier *Clf* on the training data to predict the speaker, y.

4. Obtain the predictions of *Clf* on the test set, y_{pred}, and the associated prediction probabilities, y_{pred_probs}.

5. Extract the test instances $(X_{cand}, y_{cand}) \subseteq (X_{test}, y_{pred})$ that have a prediction probability greater than some threshold, $y_{pred\_probs} \geq T$.

6. Add these to the initial train set to obtain the train and test sets for the next round $(X_{train}, y_{train}) \leftarrow (X_{cand}, y_{cand}) \cup (X_{train\_init}, y_{train\_init})$; $(X_{test}, y_{test}) \leftarrow (X, y) \setminus (X_{train}, y_{train})$.

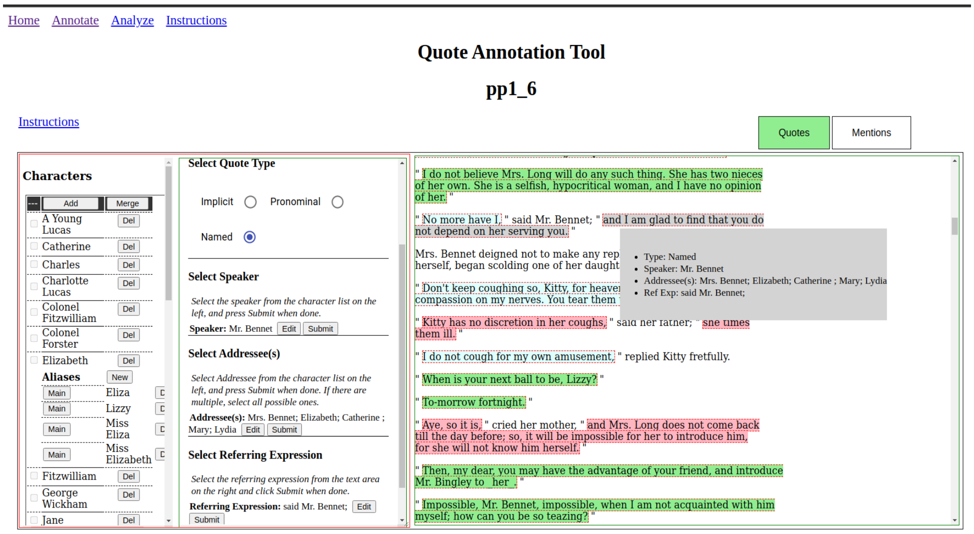

Figure 2: A screenshot of our annotation platform. The difference colors indicate the type of quotation.

7. Repeat the process from Step 2; break when there is no improvement in test performance for three consecutive iterations, or we hit 20 iterations.

Test instances that have been added to the initial train set in one round can be removed in a subsequent round if they do not satisfy the probability threshold.