Baselines for Detection and Classification of Discourse Presentation in English Narrative

Reinaldo Di Polo Mustafa Ocal Mark Finlayson
Florida International University
CASE Building, Room 265
11200 SW 8th Street, Miami, FL 33199
{rdipo001, mocal, markaf}@fiu.edu

Abstract

Discourse presentation is when speech, writing, or thought (SW&T) attributed to a discourse entity (such as a character in a narrative) is presented within a discourse. Discourse presentations can be generally broken into *direct* or *indirect*: direct presentation is when the text quotes the words or thoughts verbatim, whereas in indirect presentation the text expresses the SW&T in the narrator's or writer's own words. Automatically detecting and categorizing discourse presentations supports discourse and narrative analysis and improves attribution for downstream NLP tasks, but detecting indirect discourse presentations remains challenging due to diverse surface forms and subtle perspective shifts. We study detection and categorization of discourse presentations on a corrected version of the Semino & Short's English Narrative SW&TP corpus. We cast the task as five-way clause classification: Direct Speech & Writing, Direct Thought, Indirect Speech & Writing, Indirect Thought, and Narrative (i.e., no discourse presentation). We compare four approaches: (1) CNN; (2) generative baseline (Claude Sonnet 4.6); (3) untuned BERT, and (4) fine-tuned BERT. The CNN baseline achieves 0.43 F_1 and exhibits substantial confusion with the Narrative class. Claude achieves 0.71 F_1 but performs unevenly across classes and fails to recover Indirect Thought. BERT achieves 0.81 F_1 overall but struggles on indirect categories. The fine-tuning BERT yields strong performance (0.88 F_1), with remaining errors concentrated in Indirect Speech & Writing (F_1 = 0.60). We release our code and the corrected dataset to support reproducibility. To our knowledge, this is the first time computational approaches have been evaluated across the full range of SW&TP discourse presentation types.

1 Introduction

Discourse presentation is when a narrator or speaker conveys another agent's speech, thought, or writing (Leech and Short, 2007). Automatically identifying discourse presentation in written text is valuable for discourse analysis and downstream NLP applications that depend on correct attribution (e.g., stance, opinion mining, narrative analysis, and quotation extraction; Shibata, 2021). However, discourse presentation is difficult to model because its surface forms vary widely, from explicit quotation marks and reporting verbs to subtle forms such as indirect and free indirect speech, which often blur the boundaries between narrator and character voice (Coulmas, 1986).

Leech and Short (1981) provided the most influential account of how narratives represent a character's voice, noting distinctions such as *direct speech*, *indirect speech*, and *free indirect speech* that have since become standard. Their categories offer a principled way to describe shifts in narrative perspective and the degree to which a narrator mediates a character's utterances. Semino and Short (2004) operationalized these distinctions in the so-called SW&TP (Speech, Writing & Thought Presentation) corpus by proposing a finer-grained tagset for speech, thought, and writing presentation (25 tags), supporting the empirical study of discourse presentation phenomena. Their annotation scheme identifies, at a minimum, the spans corresponding to the presented discourse and, when overtly realized, the associated *cue* (e.g., a reporting verb) and *source* (the speaker/thinker/writer). While the SW&TP corpus is a rich resource for narratological and linguistic research, and has been influential in the study of stylistics in corpus linguistics, it surprisingly has not been previously used to test the full automatic detection of discourse presentation. In particular, we are not aware of prior work that trains supervised models to predict SW&TP discourse-presentation tags using the data in the SW&TP corpus.

In this paper, we formulate discourse presentation detection as five-way clause classification

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 1–11

task that preserves key distinctions while remaining learnable from limited supervised data. Specifically, we predict: **Direct Speech & Writing**, **Direct Thought**, **Indirect Speech & Writing**, **Indirect Thought**, and **Narrative** (which covers all other tags). We compare four approaches. First, we tested a lightweight CNN sentence encoder trained with class reweighting to address label imbalance. Second, we evaluated a generative approach by prompting a commercial large language model (LLM; Claude Sonnet 4.6) that assigns one of the five labels without task-specific fine-tuning. Third, we tested untuned BERT to classifying clauses directly. Fourth, we evaluated fine-tuned a pretrained BERT model for sequence classification using a class-weighted loss, yielding substantially stronger performance. Most of the remaining errors are concentrated in subtle indirect categories.

In the course of this work we identified issues within SW&TP that reduce its usability for computational modeling, including duplicated texts and missing or inconsistent labels. We corrected these errors using semi-automatic error detection and correction workflow followed by manual verification.

There are three major contributions of this work. First, we introduce a practical five-class discourse presentation detection formulation that collapses the Semino & Short 25-tag tagset along direct vs. indirect and speech/writing vs. thought dimensions. Second, we developed a corrected and more organized version of SW&TP for reliable supervised training and evaluation. Third, we evaluated four approaches, showing that detecting direct SW&T is relatively robust no matter the approach, but that detection of indirect SW&T still presents challenges even to our best-performing model.

The paper is organized as follows. First, we review background and prior work on discourse presentation and its automatic detection, as well as the SW&TP corpus and its underlying theory (§2). Next, we describe our method for correcting the dataset (§3). We then present our four baseline approaches (§4) and evaluate the results (§5). We conclude with a discussion of the results and propose future work (§6), and provide a summary of the contributions (§7). To support reproducibility and follow-up work, we publicly release our implementation and the corrected SW&TP-derived dataset used in our experiments[1]

[1]https://https://doi.org/10.34703/gzx1-9v95/B8NPYV

2 Related Work

Discourse presentation refers to the ways texts represent communicative and cognitive events (speech, writing, or thought) attributable to an agent other than the current narrator. Rather than a binary distinction, it encompasses a continuum: a text may quote verbatim, paraphrase, summarize that an act occurred, or blend voices so that narrator and character perspective become indeterminate (Coulmas, 1986; Fludernik, 1993). The phenomenon spans fictional narrative, autobiography, journalism, and academic prose, making it relevant to literary stylistics, discourse analysis, and computational attribution alike (Caldas-Coulthard, 1994; Waugh, 1995).

Existing work on discourse presentation can be organized into two broad traditions. First, work in the linguistics tradition—including stylistics, corpus linguistics, and narratology—have provided both theoretical perspectives and annotated data. Second, work in computational linguistics and natural language processing has addressed extraction of quotes and speaker attribution in news text. We treat each of these areas in turn.

2.1 Discourse Presentation in Linguistics and Narratology

A long tradition in stylistics distinguishes between *direct* and *indirect* presentation, as well as intermediate forms such as *free indirect* style, which blends narrator voice with a character's perspective (Leech and Short, 1981). These distinctions are central to modeling narrative voice because they capture how explicitly a text signals quotation and attribution and how strongly the narrator mediates the represented content (Norrick, 2016).

(Leech and Short, 1981) formalized many of these insights in their seminal SW&TP framework, in which they described instances of presented discourse as able to be described in terms of (i) an attributive *source* (who is speaking/thinking/writing), (ii) a *cue* (lexical or structural triggers such as reporting verbs), and (iii) the *content* being presented. In practice, however, these components are often not overtly realized: free indirect forms may lack explicit reporting clauses, sources can be implicit, and content boundaries can be ambiguous.

Most theoretical accounts that followed Leech & Short generally converge on four shared dimensions: (i) **communicative mode**, whether the content is speech, writing, or thought, each carrying different epistemic conditions for faithful represen-

tation (Fludernik, 1993; Coulmas, 1986); (ii) **degree of narrator mediation**, ranging from unfiltered direct quotation to bare assertion that a communicative event occurred, treated by most frameworks as a cline rather than discrete categories (McHale, 1978; Fludernik, 1993); (iii) **linguistic realization**, including direct quotation with reporting clauses, subordinated indirect constructions (*She said that she was tired*), free indirect style with no reporting syntax but shifted deixis and tense (*She was tired—why wouldn't they just leave?*), and narrator reports naming an act without reproducing content; and (iv) **attribution structure**, an underlying triad of source, cue, and content that is frequently only partially realized, with absent reporting verbs, implicit sources, and unclear content boundaries creating the primary challenges for both annotation and automatic detection (Coulmas, 1986; Fludernik, 1993; Norrick, 2016).

A central theoretical insight, associated with McHale (1978) and elaborated by Fludernik (1993) among others, is that presentation categories shade into one another rather than forming discrete classes (the *cline hypothesis*). Empirical corpus annotation supports this view, in that when one is exhaustively tagging naturally occurring text, a substantial share of instances cannot be easily assigned to a single class, with ambiguity concentrated at the narration/free-indirect and indirect/free-indirect boundaries (Semino et al., 1997). Notably, the direct/free-indirect boundary is less fuzzy, anchored by typographic conventions such as quotation marks (Semino et al., 1997). These properties translate into specific computational challenges. For example, since cues are the most reliable surface signal, their absence in free indirect forms makes this class especially difficult to detect (Fludernik, 1993; McHale, 1978). Detection is made even more difficult when content boundaries span multiple sentences or are found inside of nested narrative level (i.e., embedded storytelling). In journalism in particular, brief direct quotations are commonly integrated within indirect report strings (Waugh, 1995; Caldas-Coulthard, 1994), which makes the indirect presentations harder to detect.

Semino and Short operationalized Leech and Short's theoretical categories in the SW&TP corpus by providing a fine-grained annotation scheme spanning *speech*, *thought*, and *writing* (Semino and Short, 2004). In total, the scheme comprises 25 tags: 8 speech tags, 8 writing tags, 8 thought tags, and a single narrative tag (N). At the level of communicative mode, the scheme includes tags for *direct* realizations (e.g., DS Direct Speech, FDS Free Direct Speech; and their writing counterparts DW/FDW) and *indirect* realizations (e.g., IS Indirect Speech, FIS Free Indirect Speech; and IW/FIW for writing). Analogous tags capture the presentation of thought (e.g., DT/FDT for direct thought and IT/FIT for indirect thought). In addition to these content-bearing categories, SW&TP includes narrator-oriented tags that report communicative activity without rendering the utterance/thought/writing itself (e.g., narrator representations of speech/thought/writing acts such as NRS, NRSA, NRT, NRW, etc.), as well as related narrative categories such as internal narration.

The corpus contains 120 texts of approximately 2,000 words each, totaling 258,348 words of late twentieth-century written British English across 14,380 sentences. It is balanced across three narrative genres, with 40 samples per genre: prose fiction (87,709 words), news reports (83,603 words), and autobiography (87,036 words). Each genre is further subdivided into "serious" and "popular" sections (e.g., broadsheet vs. tabloid press; highbrow vs. popular fiction). The prose fiction and (auto)biography sections include further narrative-voice stratification: fiction is subdivided into first- vs. third-person narration. The (auto)biography section aligns naturally with this distinction (biographies are third-person; autobiographies first-person). Crucially, the entire SW&TP corpus is manually annotated for all the discourse presentation categories.

2.2 Computational Work on Discourse Presentation

We are not aware of prior work that *directly* trains and evaluates supervised models to predict the full set of SW&TP tags in the SW&TP corpus itself. Instead, related research can be grouped into three closely connected strands that address overlapping phenomena under different label inventories and task formulations.

Quotation and reported-speech detection in news (direct + indirect) A substantial line of NLP work seeks to detect direct speech (a.k.a., *reported speech*) in newswire by detecting quotation spans and/or attributing them to speakers. Early systems typically combined linguistic cues with rule-based or feature-driven models to identify direct and indirect reported speech (and related

evidential constructions), reporting high precision (e.g., 99%) but lower recall (e.g., 74%) for quotation span detection on newswire text (Krestel et al., 2008). Subsequent work broadened the scope beyond direct quotations, explicitly addressing *indirect* and *mixed* quotations and jointly modeling quotation extraction and attribution (Pareti et al., 2013). More recent resources and models have focused on extracting and attributing quotation content, cues, and sources, often with modern neural architectures and richer annotation of quotation structure, reporting an overall F_1 of 0.71 for quotation detection, with direct quotations generally easier (e.g., F_1 around 0.91) than indirect quotations (e.g., F_1 around 0.65) (Papay and Padó, 2019). While these tasks strongly overlap with SW&TP categories such as direct vs. indirect speech/writing, they typically do not aim to reproduce the full SW&TP tagset, and they are often optimized for journalistic quotation conventions rather than the stylistic variety present in narrative prose.

Computational work on narrative fiction (direct speech and attribution; FID-adjacent phenomena) A second strand of work focuses on literary narrative, where reported speech detection is often used as a component for higher-level analyses such as modeling the interactions of character. For example, prior work extracts and attributes quoted speech in novels to support social-network construction from fiction (Elson et al., 2010). Using character-trigrams, they achieved a 0.67 F_1 score on speech adjacency detection, and a 0.41 F_1 score on spoken mention detection (Elson et al., 2010). Related literary-oriented quotation research emphasizes robust quotation detection and attribution under stylistic variation (e.g., dialogue formatting and punctuation idiosyncrasies), and provides corpora and models aimed at literary text rather than news (Muzny et al., 2017; Papay and Padó, 2020). Although these efforts overlap with SW&TP categories—especially direct discourse categories—they generally do not model the broader SW&TP inventory spanning speech, writing, and thought presentation, and free indirect forms are often treated indirectly (e.g., as difficult edge cases) rather than as explicit target labels.

SW&TP tagging in other languages and DH-oriented corpora Finally, digital humanities and corpus-linguistic projects have developed SW&TP-inspired resources and taggers in languages other than English. For example, Brunner reports early experiments in German on automatically recognizing and classifying a tagset similar to Leech & Short's SW&TP tagset using surface cues and computational methods (Brunner, 2013), achieving F_1 of 0.87 for *direct*, 0.71 for *indirect*, 0.40 for *free indirect*, and 0.58 for *reported* speech, writing, and thought. More recently, the German REDEWIEDERGABE corpus provides detailed annotations across a large corpus and is explicitly designed to support machine learning (Brunner et al., 2020a). Associated taggers explore modern embeddings (including BERT-style representations) and report differing difficulty levels across direct, indirect, and free indirect categories (Brunner et al., 2020b). Their best models achieve an F_1 of 0.85 for *direct*, 0.76 for *indirect*, 0.59 for *free indirect*, and 0.60 for *reported* speech, writing, and thought. These results reinforce a consistent trend across settings: indirect and especially free-indirect categories are harder to recognize than direct ones, even with modern embeddings.

Taken together, prior work provides suggests that direct discourse presentation can be easily modeled, but it leaves a gap for SW&TP-style discourse-presentation categories on English narrative data. Our work addresses this gap by adopting a practical label collapse grounded in SW&TP distinctions and establishing a reproducible baseline on a cleaned SW&TP-derived dataset.

3 Corpus Corrections

The original SW&T corpus comprised 14,380 sentence-level annotations, each labeled with Speech, Writing, and Thought (SW&T) presentation tags following the Semino and Short (2004) annotation scheme. A systematic audit revealed two principal quality issues that necessitated correction: (1) missing or incorrect source arguments in attributed relations, and (2) sentences carrying multiple SW&T tags without a one-to-one correspondence between spans and labels. We describe the hybrid human–machine methodology employed to address each issue in turn.

First, we identified 198 duplicate sentences in the original corpus. Removing these resulting in a deduplicated set of 14,182 unique source sentences prior to span expansion.

Next, a non-trivial portion of the original annotations contained empty source tags, leaving cue verbs (e.g., *said*, *claimed*) without an attributed agent. To recover these missing sources automati-

cally, we employed a hybrid approach combining AllenNLP Semantic Role Labeling (SRL; Gardner et al., 2018) with few-shot prompting via GPT-4o. The SRL module identified ARG0 and ARG1 candidates for each predicate in the sentence, providing structured argument candidates; these were then passed to GPT-4o with few-shot demonstrations to determine the most plausible source span for each cue, taking into account syntactic context and discourse structure.

All automatically predicted source corrections were subsequently reviewed by a team of five trained annotators. Each annotator independently assessed whether the predicted source was correct, marking it as either *Correct* or *Incorrect*; in cases where the predicted source was deemed incorrect, annotators proposed the correct source span. Manual corrections were accepted only when all five annotators reached unanimous agreement, ensuring high-confidence revisions. Annotators were specifically instructed to treat passive constructions (e.g., *It was hoped that…*, *It was reported that…*) as having an empty source when the originating agent is not explicitly specified in the text, rather than defaulting to the grammatical subject.

Finally, the original annotation scheme permitted a single sentence to carry multiple SW&T labels, resulting in compound tags (e.g., `IS-FIS`, `NRS-NRSAP`) that conflated distinct presentational modes within a single span. To impose a stricter one-label-per-span constraint, we segmented sentences at clause boundaries so that each resulting phrase received exactly one SW&T tag. Sentence-internal phrases that could not be unambiguously assigned through rule-based decomposition were flagged and submitted to the full annotator panel. Following the Semino and Short (2004) guidelines, all five annotators independently labeled each flagged phrase, and the majority label was adopted. This process expanded the annotation set from 14,182 sentence-level instances to 24,519 span-level instances. Table 1 summarizes the class distribution and average clause length across the resulting 24,519 span-level instances.

Class	Train	Test	Total	Avg. len.
DS&W	4,270	474	4,744	8.5
DT	1,145	127	1,272	7.4
IS&W	1,463	163	1,626	12.4
IT	1,143	127	1,270	12.5
N	14,046	1,561	15,607	11.3
Total	**22,067**	**2,452**	**24,519**	**10.7**

Table 1: Dataset statistics for the corrected SW&TP-derived corpus after span decomposition. Train/Test split is stratified 90/10. *Avg. len.* is the mean clause length in whitespace-tokenized tokens. DS&W = Direct Speech & Writing; DT = Direct Thought; IS&W = Indirect Speech & Writing; IT = Indirect Thought; N = Narrative.

4 Methods

We evaluated four supervised modeling approaches discourse presentation detection on the corrected SW&TP dataset: CNN, Generative, Untuned BERT, and Fine-tuned BERT. All approaches shared the same task formulation, data preparation, and preprocessing. We formulate reported-speech detection on SW&TP as a clause-level, five-class classification problem. Each clause is assigned exactly one label from the following set: **Direct Speech & Writing** (DS, FDS, DW, FDW), **Direct Thought** (DT, FDT), **Indirect Speech & Writing** (IS, FIS, IW, FIW), **Indirect Thought** (IT, FIT), and **Narrative** (all remaining SW&TP tags). Table 2 provides the full mapping from the original SW&TP tags to the five classes used throughout this paper. This collapse preserves the core dimensions emphasized in the SW&TP framework (namely, direct vsindirect) as well as preserving the distinction between speech/writing and thought, which are qualitatively different (in that speech and writing have observable instantiations in words or text, while thought does not). For all approaches, we removed HTML markup and normalized whitespace. For the CNN baseline, we remove HTML markup, replace non-alphabetic characters with whitespace, delete single-character tokens, and collapse repeated spaces.

4.1 CNN

We convert each sentence into a sequence of subword token identifiers using the `bert-base-uncased` tokenizer (Devlin et al., 2018). This step provides a robust, vocabulary-controlled representation that handles morphological variation and rare words via WordPiece-style segmentation (Wu et al., 2016). To obtain fixed-size inputs for batching and convolutional encoding, we zero-pad or truncate each sequence to a maximum length of 256 tokens (`MAX_LEN` $= 256$). During training, we use a batch size of 32.

The encoder itself is a lightweight convolutional neural network (Fukushima, 1980) for sentence

Five-class label	Abbrev.	Original SW&TP tags
Direct Speech & Writing	DS&W	DS, FDS, DW, FDW
Direct Thought	DT	DT, FDT
Indirect Speech & Writing	IS&W	IS, FIS, IW, FIW
Indirect Thought	IT	IT, FIT
Narrative	N	N, NRS, NRSA, NRSAT, NRT, NRTA, NRTAT, NRW, NRWA, NV, NI, and all remaining narrator-oriented tags

Table 2: Mapping from the original SW&TP annotation scheme (Semino and Short, 2004) to the five classes used in this paper. Tag abbreviations follow Semino and Short: DS = Direct Speech, FDS = Free Direct Speech, DW = Direct Writing, FDW = Free Direct Writing, IS = Indirect Speech, FIS = Free Indirect Speech, IW = Indirect Writing, FIW = Free Indirect Writing, IT = Indirect Thought, FIT = Free Indirect Thought, NRS = Narrator's Report of Speech, NRSA = Narrator's Representation of Speech Act, NRT = Narrator's Report of Thought, NRTA = Narrator's Representation of Thought Act, NV = Narrator's Report of Voice, NI = Narration of Internal States.

classification. Given an input sequence corresponding to a clause, we learn a dense vector for each token using an embedding layer with dimension 100. We then apply three parallel one-dimensional convolutional layers with 128 filters each and kernel widths 2, 3, and 4. Each convolution is followed by a ReLU nonlinearity and global max-pooling over time, yielding three pooled vectors. We concatenate these vectors and pass the result through a 128-dimensional fully connected layer with ReLU activation and dropout ($p = 0.5$). The final layer is a five-way softmax classifier.

We train the CNN baseline using the Adam optimizer with sparse categorical cross-entropy loss and optimize all parameters end-to-end. Training is performed for five epochs (`NB_EPOCHS` $= 5$) with a batch size of 32. Because the label distribution is imbalanced, we compute class weights from the training labels and apply them during optimization so that errors on minority classes receive a larger penalty. Final performance is reported on a held-out test portion (10% of the data), and we use confusion-matrix inspection to characterize error patterns.

4.2 LLM (Claude Sonnet 4.6)

We next evaluated a prompt-based large language model (LLM) baseline using *Claude Sonnet 4.6*. [2] The goal of this experiment is to assess how well a general-purpose LLM can perform SW&TP-style reported-speech classification *without* task-specific fine-tuning, relying only on natural-language instructions and examples. We prompt the model to assign exactly one label from the same five-class inventory used throughout the paper: **Direct Speech & Writing** (DS), **Direct Thought** (DT), **Indirect Speech & Writing** (IS), **Indirect Thought** (IT), and **Narrative** (N). The prompt includes concise definitions of each class and one illustrative example per label. For reproducibility, the full prompt text is provided in Appendix A.

To enable a fair comparison, we evaluate the LLM on the same held-out test split and report the same metrics used for supervised models: accuracy, micro-F_1, macro-F_1, and weighted-F_1, along with per-class precision, recall, and F_1. Each test instance consists of a single clause (the `Text` field) presented to the LLM in isolation. The LLM is instructed to return *only* one of the five label tokens (DS/DT/IS/IT/N), which simplifies parsing and avoids reliance on free-form rationales.

We post-process model outputs by mapping valid responses directly to the corresponding class. If a response does not exactly match one of the allowed labels (e.g., additional text, alternative tag names, or invalid strings), we treat it as an invalid prediction and map it to the majority **Narrative** label to avoid artificially improving performance through selective filtering.[3] We then compute corpus-level and per-class scores against the gold labels in the test set using standard implementations of precision/recall/F_1.

4.3 Untuned BERT

We next evaluated an untuned BERT classification. Each instance is lightly normalized (whitespace cleanup only) and encoded with the `bert-base-uncased` tokenizer, truncated to a maximum sequence length of 256 WordPiece tokens.

[2] Claude Sonnet 4.6 was selected because, at the time this research was conducted, it represented the best-performing generally available model in the Claude family and was among the top-performing LLMs on general language understanding benchmarks. We leave evaluation of open-weight alternatives (e.g., Llama, Gemma, Qwen) to future work.

[3] In our runs, invalid outputs were rare due to the strict output format requested in the prompt.

We then pass the tokenized inputs through a frozen BERT encoder (all parameters fixed) and represent each text by a single 768-dimensional vector derived from the model's sentence-level representation (the pooled [CLS] embedding via pooler_output, with a fallback to the [CLS] hidden state when pooling is unavailable).

On top of these fixed representations, we train a lightweight discriminative classifier without updating BERT. In the simplest variant, we fit a linear SVM with class balancing to mitigate label imbalance. This baseline isolates the contribution of contextual BERT representations from task-specific fine-tuning and provides a strong, reproducible comparison point for fully fine-tuned transformer models.

4.4 Fine-tuned BERT

For BERT fine-tuning, we use the same `bert-base-uncased` tokenizer (Devlin et al., 2018) but retain punctuation and casing normalization consistent with the model's pre-training. Each sentence is tokenized with truncation to 256 tokens (`MAX_LENGTH` $= 256$). Instead of padding all sequences to a fixed maximum in advance, we apply dynamic padding within each batch using a padding collator, producing standard Transformer inputs (e.g., `input_ids` and `attention_mask`).

We fine-tune the pre-trained BERT encoder with a sequence-classification head, using five output labels. The model is trained with cross-entropy loss, and we incorporate class imbalance by weighting the loss with class weights computed from the training split. We use a stratified 90/10 train–validation split with a fixed random seed to ensure reproducibility and to preserve class proportions across splits. Fine-tuning is carried out for three epochs with batch size 16 and learning rate 2×10^{-5}, using weight decay (0.01) and a short warmup schedule (6% of total steps).

5 Results

We report results for all four approaches on the five-class SW&TP formulation. We evaluate using accuracy and F_1 (micro, macro, and weighted), and report per-class F_1 to characterize error patterns across discourse presentation categories. Full per-class results are shown in Table 3.

5.1 CNN

The CNN baseline yields a micro-F_1 of 0.431 and macro-F_1 of 0.212, substantially below all other approaches. As Table 3 shows, performance is dominated by the Narrative class ($F_1 = 0.610$), while all four discourse presentation categories score below 0.22. The CNN tends to over-predict Narrative and struggles to reliably separate direct from indirect categories when cues are subtle or context-dependent, motivating richer contextual encoders for SW&TP prediction. The confusion matrix for this approach is provided in Appendix B.

5.2 LLM (Claude Sonnet 4.6)

The prompted LLM achieves reasonable performance, but is highly uneven across classes. Claude handles Narrative and DS&W reasonably well, suggesting that overt surface cues are mostly sufficient for these categories. However, it fails almost entirely on indirect and cognitively-oriented categories, and notably fails completely on Indirect Thought.

Overall, these findings suggest that while prompt-based LLMs can serve as a reasonably strong baseline for high-frequency classes and overt direct discourse, supervised fine-tuning remains essential for robust performance across the full five-class label space, particularly for indirect and cognitively-oriented categories such as Indirect Thought.

5.3 Untuned BERT

The untuned BERT approach achieves a micro-F_1 of 0.807 and macro-F_1 of 0.694, making it a strong feature-extraction baseline despite no task-specific fine-tuning of the encoder. As shown in Table 3, direct categories are handled well, with DS&W and DT reaching F_1 of 0.823 and 0.844 respectively. However, performance degrades substantially on indirect categories: IS&W reaches only $F_1 = 0.509$ and IT $F_1 = 0.419$, indicating that frozen BERT representations, while expressive, do not capture the subtle cues that distinguish indirect discourse from narration without further supervision.

5.4 Fine-tuned BERT

Fine-tuning BERT substantially improves performance across all metrics, achieving micro-$F_1 = 0.880$ and macro-$F_1 = 0.823$. As Table 3 shows, this approach achieves the best score in every class. Direct categories are handled with high accuracy (DS&W $F_1 = 0.862$, DT $F_1 = 0.950$), and Indirect Thought improves markedly to $F_1 = 0.781$. Indirect Speech & Writing remains the most challenging category ($F_1 = 0.603$), suggesting that

Approach	DS&W	DT	IS&W	IT	N	Macro-F_1	Micro-F_1
CNN	0.217	0.067	0.105	0.062	0.610	0.212	0.431
Claude Sonnet 4.6	0.685	0.221	0.286	0.000	0.811	0.401	0.707
BERT+SVM	0.823	0.844	0.509	0.419	0.875	0.694	0.807
Fine-tuned BERT	**0.862**	**0.950**	**0.603**	**0.781**	**0.918**	**0.823**	**0.880**

Table 3: Per-class F_1 scores and aggregate Macro- and Micro-F_1 for all four approaches on the five-class SW&TP task. DS&W = Direct Speech & Writing; DT = Direct Thought; IS&W = Indirect Speech & Writing; IT = Indirect Thought; N = Narrative. For single-label multi-class classification, Micro-F_1 equals accuracy. Bold indicates best score per column.

indirect realizations with weak or implicit cues continue to pose difficulties even for fine-tuned contextual representations. The confusion matrix for this approach is provided in Appendix B.

5.5 Comparison

Table 3 summarizes results across all four approaches and reveals two consistent patterns. First, fine-tuned BERT outperforms all other approaches on every class, suggesting that task-specific supervision is essential for robust discourse presentation detection. Second, indirect categories are consistently harder than direct across all approaches: IS&W and IT lag behind DS&W and DT in every system, with the gap largest for the CNN and LLM. Untuned BERT occupies an informative middle ground: its strong direct-category performance shows that BERT representations carry substantial signal, but its indirect-category scores reveal that fine-tuning is necessary to close the remaining gap. The LLM exhibits a qualitatively different failure pattern from the other approaches: rather than degrading gradually across indirect categories, it collapses entirely on Indirect Thought, indicating that the linguistic properties of this class are not recoverable from instructions alone.

6 Discussion and Future Work

Our experiments highlight a clear trade-off between lightweight approaches and contextualized Transformer models for five-way SW&TP classification. CNN provides an efficient starting point, but it performs substantially worse than all other approaches: in particular, it is dominated by confusion with the Narrative class, yielding a relatively poor overall performance. Claude Sonnet 4.6 occupies a qualitatively different position, achieving reasonable performance on high-frequency and overtly-cued categories (Narrative, DS&W) but failing entirely on Indirect Thought, revealing that natural-language instructions alone are insufficient in this case to recover subtle cognitively-oriented categories. Untuned BERT occupies an informative middle ground, with strong performance on direct categories (DS&W) demonstrating that BERT embeddings carry substantial signal for discourse presentation even without fine-tuning. However, it struggles on indirect categories (IS&W, IT), suggesting that capturing the subtle cues that distinguish indirect discourse from narration requires task-specific adaptation of the encoder itself. In contrast, fine-tuned BERT achieves strong overall performance and substantially improves recognition of Direct Speech & Writing, Direct Thought, Indirect Thought, and Narrative. These gains suggest that contextualized representations are crucial for modeling discourse-presentation cues that go beyond local n-gram patterns and that depend on syntactic and semantic context.

Despite the improvement, the results also expose persistent challenges. Indirect Speech & Writing remains the most difficult category even under the best-performing model (Fine-tuned BERT), indicating that indirect realizations, often characterized by weaker typographic cues and greater reliance on reporting constructions or discourse context, are still frequently confused with neighboring categories, especially Narrative. This pattern is consistent with the underlying SW&TP distinctions: indirect and free-indirect forms can be signaled implicitly, can span multiple clauses, and can blur narrator voice with a character's perspective. Notably, the LLM's failure on Indirect Thought suggests that this collapse is not merely a data-sparsity problem but reflects a genuine difficulty in recovering implicit cognitive attribution from instructions alone. Moreover, although fine-tuned BERT reduces these confusions considerably relative to both CNN and Untuned BERT, the remaining errors suggest that clause-level classification alone cannot capture all cues needed for robust indirect discourse detection.

These observations point to several avenues for

future work. First, one can incorporate broader context by conditioning predictions on neighboring sentences or paragraph windows, using hierarchical encoders or sliding-window Transformers to model cross-sentence dependencies (e.g., attribution established earlier, or multi-sentence quotation continuity). Second, rather than predicting a single label per clause, one could use span-based modeling that explicitly identifies the source, cue, and content components of discourse presentation, enabling extraction and attribution rather than only categorization. Third, while collapsing the SW&TP tagset into five classes improves learnability, it obscures distinctions important for fine-grained stylistic analysis; future work should therefore evaluate how well models scale to a richer label inventory and whether multi-task learning can jointly support coarse and fine-grained predictions. Finally, it would be useful to examine genre-specific performance differences (i.e., across prose fiction, news, and (auto)biography), and conduct in-depth, targeted error analyses and data augmentation for the most ambiguous categories (notably indirect and free-indirect forms), to better understand where models generalize and where they fail.

7 Contributions

This is the first study of which we are aware that trains and evaluates supervised models for automatic prediction of SW&TP-style discourse-presentation categories on the *English* SW&TP corpus. Beyond this, the paper makes three contributions to discourse presentation detection and classification. First, we introduce a five-class task formulation that collapses the original SW&TP tagset along direct vs. indirect and speech & writing vs. thought dimensions, yielding a learnable label space that still reflects central narratological distinctions. Second, we constructed and release a cleaned, more organized SW&TP-derived dataset tailored for supervised learning: we remove duplicate/repeated texts, resolve missing tags, map each instance to one of the five target labels, and provide a clause-level representation that supports straightforward modeling and evaluation. Third, we provide four approaches trained and tested with these data: (1) a CNN model that establishes a lightweight reference point; (2) a prompt-based LLM baseline (Claude Sonnet 4.6) that shows limits of few-shot inference for indirect categories; (3) an Untuned BERT encoder; and (4) a fine-tuned BERT that substantially improves performance and clarifies which categories remain challenging.

References

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and linguistic computing*, 28(4):563–575.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020a. Corpus redewiedergabe. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 803–812.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020b. To bert or not to bert-comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *SwissText/KONVENS*.

Carmen Rosa Caldas-Coulthard. 1994. On reporting reporting: The representation of speech in factual and factional narratives. In Malcolm Coulthard, editor, *Advances in Written Text Analysis*, pages 295–308. Routledge, London.

Florian Coulmas. 1986. Reported speech: Some general issues. *Direct and indirect speech*, 31:1–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Monika Fludernik. 1993. *The Fictions of Language and the Languages of Fiction*. Routledge, London.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew E Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of workshop for NLP open source software (NLP-OSS)*, pages 1–6.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Geoffrey N. Leech and Mick Short. 1981. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Longman, London.

Geoffrey N Leech and Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose*. 13. Pearson Education.

Brian McHale. 1978. Free indirect discourse: A survey of recent accounts. *Poetics and Theory of Literature*, 3:249–287.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 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.

Neal R Norrick. 2016. Indirect reports, quotation and narrative. In *Indirect reports and pragmatics: Interdisciplinary studies*, pages 93–113. Springer.

Sean Papay and Sebastian Padó. 2019. Quotation detection and classification with a corpus-agnostic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 888–894, Varna, Bulgaria. INCOMA Ltd.

Sean Papay and Sebastian Padó. 2020. Riqua: A corpus of rich quotation annotation for english literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 835–841.

Silvia Pareti, Tim O'keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999.

Elena Semino and Mick Short. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. Routledge.

Elena Semino, Mick Short, and Jonathan Culpeper. 1997. Using a corpus to test a model of speech and thought presentation. *Poetics*, 25(1):17–43.

Masaki Shibata. 2021. Reported speech as persuasion: A discourse analysis of japanese journalism. *Japanese Studies*, 41(2):221–239.

Linda R. Waugh. 1995. Reported speech in journalistic discourse: The relation of function and text. *Text*, 15(1):129–173.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, and Hideto Kazawa. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A Claude Sonnet 4.6 Prompting Details

This appendix provides the full prompt used for the Claude Sonnet 4.6 generative baseline (Section 4.2). Each test sentence was submitted individually as the value of {SENTENCE} below; no conversation history or prior labels were passed across instances.

You are a linguistic annotation assistant. Your task is to assign ONE label to the given sentence based on the SW&TP-inspired five-class scheme below. Return ONLY the label (one of: DS, DT, IS, IT, N). Do not output anything else.

LABELS (choose exactly one):

1) DS — Direct Speech & Writing
Use DS when the sentence presents speech or writing in a direct form, typically with quotation marks or direct quotation style, OR direct written communication presented as if quoted. Includes: direct speech/writing and free direct speech/writing.
Examples:

- "I can't believe it," she said.
- Dear John, I am leaving tomorrow.

2) DT — Direct Thought
Use DT when the sentence presents a character's thought directly (often first-person, present-tense, sometimes italicized or exclamatory), including free direct thought.
Example:

- What am I going to do now?

3) IS — Indirect Speech & Writing
Use IS when the sentence reports speech or writing indirectly (typically with that/if/whether clauses, reported questions, or paraphrases), including free indirect speech/writing.
Examples:

- She said that she couldn't believe it.
- The letter stated that the meeting was cancelled.

4) IT — Indirect Thought
Use IT when the sentence reports a thought indirectly (paraphrased or embedded under thinking/knowing/wondering), including free indirect thought.
Example:

- He wondered whether he had made the right choice.

5) N — Narrative (everything else)
Use N when the sentence is plain narration, description, exposition, action, scene-setting, or narrator commentary that is NOT presenting speech/thought/writing content as DS/DT/IS/IT.
Example:

- The rain fell steadily over the empty street.

DECISION RULES:

- If it is clearly quoted/direct → prefer DS (speech/writing) or DT (thought).
- If it is clearly reported/paraphrased (said/thought/wrote + that/if/whether, etc.) → prefer IS/IT.
- If there is no speech/thought/writing presentation → N.

Sentence: {SENTENCE}

B Confusion Matrices

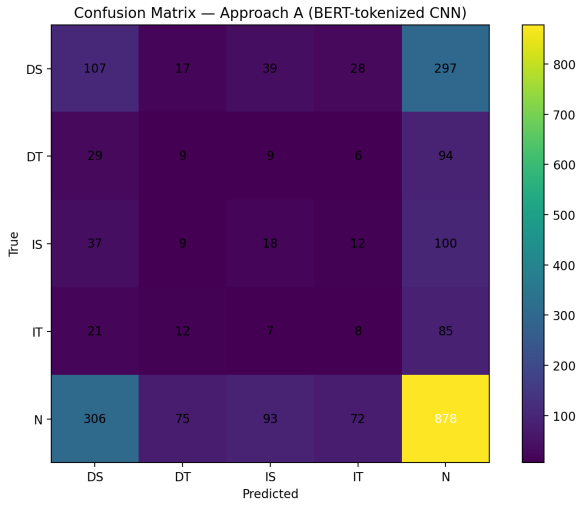

Figure 1: Confusion matrix for BERT-tokenized CNN baseline on the five-class SW&TP task.

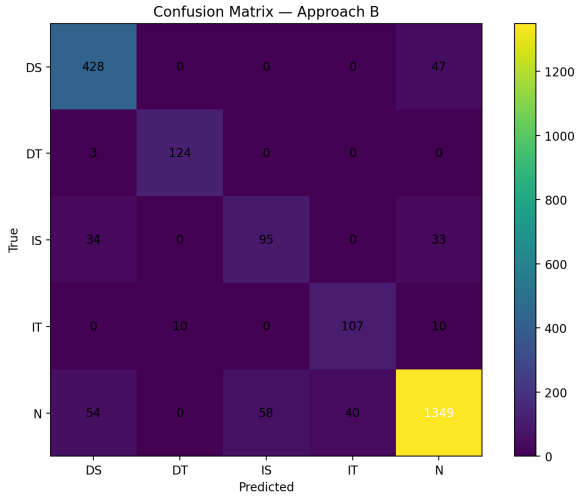

Figure 2: Confusion matrix for Fine-tuned BERT. Rows are gold labels and columns are model predictions.