



Annotating Free Indirect Discourse in Victorian Fiction With a Shared Tagging Scheme

Andre Dang^{1*}, Jack Khumalo², Tomasz Ristic³

¹Department of Social Sciences, Fjordland Graduate Institute of Social Sciences, Chile

²Department of Social Sciences, Mekong University Centre for Social Sciences, Malaysia

³Department of Social Sciences, Fjordland Institute of Social Sciences, Hungary

Article Info

Volume 1, Issue 1, 2026

[doi: 10.62715/ijesl.2026.v1.i1.pp19-24](https://doi.org/10.62715/ijesl.2026.v1.i1.pp19-24)

Received: 14 September 2025

Accepted: 16 November 2025

Published: 19 January 2026

Abstract

Free indirect discourse, in which a narrator's sentence carries a character's words or thoughts without quotation marks or a reporting clause, is central to accounts of the Victorian novel and notoriously hard to pin down. Critics disagree about individual passages, computational methods need labelled examples that are scarce for nineteenth-century English fiction, and projects that annotate discourse presentation use categories that are only partly compatible. This article proposes a shared tagging scheme and workflow intended to let separate teams produce comparable annotations. The scheme marks spans of presented discourse on four independent layers: mode (speech, thought or writing), form (direct, indirect, free indirect or narrator's report), attribution and certainty. It supplies a list of linguistic cues with an ordered decision procedure, treats ambiguity between narrator and character as a value to be recorded and not an error to be removed, and maps onto the Text Encoding Initiative guidelines. The workflow covers piloting, written guidelines, double annotation, agreement measurement and adjudication. The scheme is a proposal derived from the narratological and annotation literature. It has not been applied to a corpus or tested for reliability here, and the conditions for such a test are described.

Keywords: *Free indirect discourse, Literary annotation, Victorian novel, Inter-annotator agreement, Text Encoding Initiative, Narratology*

© 2026 Andre Dang et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A sentence such as 'She would not, surely, be expected to give way now', an example constructed for this article, shows the phenomenon in small. Its third person and past tense belong to a narrator. Its 'surely', its 'now' and its note of indignation belong to a character. This is free indirect discourse. It has been described as the characteristic technique of the nineteenth-century novel for rendering minds and voices (Pascal, 1977; Cohn, 1978), and arguments about irony, sympathy and narrative authority in Austen, Eliot, Dickens and Trollope regularly turn on whether a given sentence is the narrator's judgement or a character's.

* Corresponding author: Andre Dang, Department of Social Sciences, Fjordland Graduate Institute of Social Sciences, Chile. E-mail: andre.dang@fjordland.edu

© 2026 Andre Dang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Large-scale study of the technique is held back by a practical gap. Automatic recognition of direct speech is comparatively easy because quotation marks do most of the work. Free indirect discourse has no such marker, and systems that attempt to find it need hand-labelled examples for training and testing (Brunner, 2013). The most thorough annotated corpus of discourse presentation in English covers twentieth-century writing (Semino and Short, 2004). For Victorian fiction, such annotation as exists has been produced for the purposes of individual projects, with category systems that differ in detail, so that results cannot easily be pooled or compared.

This article proposes a tagging scheme and an accompanying workflow designed to be shared. The aim is modest: a small set of categories, defined closely enough that trained annotators in different places would be likely to apply them alike, and an explicit way of recording the cases where they would not. The scheme is presented as a design. It has not been applied to a body of texts for this article, and no agreement figures are reported.

2. Background

Two strands of earlier work inform the design: the narratological debate about what free indirect discourse is, and the experience of projects that have annotated discourse presentation.

2.1. Accounts of free indirect discourse

The theoretical literature offers two main pictures. In the dual-voice account, associated with Pascal (1977), the narrator's voice and the character's are both present in the sentence, and the reader's sense of irony or sympathy arises from their combination. Banfield (1982) rejected this on grammatical grounds: sentences of represented speech and thought have, in her analysis, a single point of view, that of the character, and no speaker at all. McHale (1978) had already shown how varied the phenomenon is and how poorly it is captured by a purely syntactic derivation from direct or indirect speech, and Fludernik (1993) argued at length that free indirect discourse is recognised by interpretation in context, with linguistic features serving as prompts and not as proof. A reading experiment by Bray (2007) put the dual-voice hypothesis to actual readers and found reason to question it. For annotation the lesson is that the category has a clear core and a wide, disputed margin, and a scheme has to provide for both.

2.2. Existing annotation models

Leech and Short (2007) arranged the forms of speech and thought presentation on a scale of narratorial control, from the narrator's bare report that speech occurred to free direct discourse with no narratorial framing. Semino and Short (2004) tested and extended that model by annotating a corpus of fiction, news reports and biography, adding a parallel scale for writing presentation and several new categories, and they documented the ambiguities their annotators met. For German, Brunner (2013) annotated narrative texts for direct, indirect, free indirect and reported forms and used the result to develop and compare automatic recognisers, with free indirect discourse the most difficult form to detect. Beyond discourse presentation, Gius and Jacke (2017) examined why annotators of narrative categories disagree and separated simple mistakes and gaps in the guidelines from disagreements that arise because the text itself supports more than one analysis. The scheme proposed below borrows its categories from the first two sources and its attitude to disagreement from the last.

3. Why Victorian Fiction Is a Hard Case

Victorian novels present particular difficulties. Their narrators are talkative and opinionated, given to general reflections in the present tense, so that an evaluative sentence may be the narrator's wisdom, the character's self-justification, or the first mimicking the second. Gunn (2004), writing on Austen, argued that free indirect discourse is best understood as the narrator's imitation of a character within the narrator's own discourse, a view that fits Eliot and Trollope well and that makes a sharp boundary hard to draw. Dickens uses the form for collective and satirical voices, the opinions of a dinner table or a government office, which have no single source. *Bleak House* is partly narrated in the present tense, which removes the backshifted tense on which many definitions rely. Dialect and long sentences with several

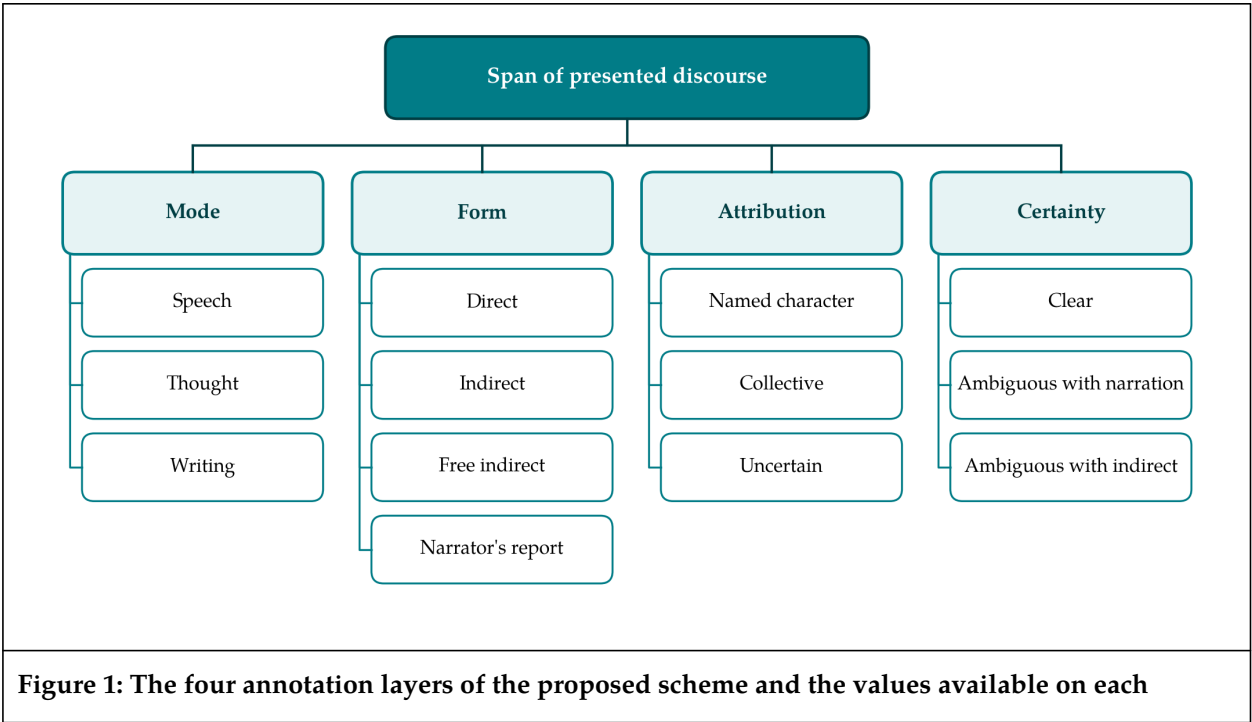
embedded viewpoints add further complications. A scheme tuned to twentieth-century prose will not transfer without adjustment.

4. The Proposed Scheme

The scheme rests on one decision: the properties of a passage that annotators are asked to judge are separated into layers that can be decided, and disagreed about, independently. Figure 1 shows the four layers and their values.

4.1. Units and layers

The unit of annotation is a continuous span of text that presents the speech, thought or writing of a character or group. A span may be shorter than a sentence or run across several. Each span receives one value on each layer. Mode records whether what is presented is speech, thought or writing. Form records whether it is direct, indirect, free indirect or a narrator's report of a discourse act, a simplification of the fuller scale in [Semino and Short \(2004\)](#) chosen for the sake of reliable application. Attribution names the source: an identified character, a collective, or uncertain. Certainty records how confident the annotator is in the form value. Keeping mode separate from form avoids a common difficulty in Victorian prose, where an annotator may be sure that a passage is free indirect but unsure whether it is spoken or thought.



4.2. Cues and decision procedure

Annotators need to know what counts as evidence. Table 1 lists six cues taken from the linguistic descriptions of the form ([Banfield, 1982](#); [Fludernik, 1993](#)), with the reason why none is sufficient alone. The decision procedure is ordered. The annotator first asks whether the span presents discourse at all, as opposed to narration of events or description. If it does, quotation marks or an explicit reporting clause with a subordinated proposition settle the form as direct or indirect. If neither is present, the annotator checks the cues. A span is tagged free indirect when it keeps the person and tense of the surrounding narration, shows at least one further cue, and can be attributed in context to a character's act of speaking, thinking or writing. Where the cues are present but attribution to the narrator remains equally plausible, the span is still tagged, with the certainty layer carrying the doubt.

4.3. Recording ambiguity

The certainty layer has three values: clear, ambiguous between free indirect discourse and narration, and

ambiguous between free indirect and indirect discourse. This is the scheme's main departure from practice that forces a single label. Much of the critical interest of the technique lies in sentences that cannot be assigned, and a resource that silently resolves them misrepresents the texts. Recording ambiguity as a value has a second benefit. It allows agreement to be calculated separately for the clear cases, where high reliability can reasonably be demanded, and for the margin, where disagreement between competent readers is informative. An optional free-text note lets annotators state the competing readings.

Table 1: Cues for identifying free indirect discourse and the reason each is not decisive alone

Cue	What to look for	Why it is not decisive
Person and tense	Third person and narrative past where direct speech would shift them	Identical to plain narration; lost in present-tense narration
Deixis	Now, here, tomorrow, this, anchored to the character	Narrators also use vivid deixis
Expressive syntax	Questions, exclamations and fragments with no reporting clause	May be the narrator's own rhetoric
Modality	Would, must, might, surely, expressing a character's judgement	Sententious narrators generalise in similar terms
Idiom and register	Vocabulary typical of one character or social group	Requires knowledge of the whole novel
Framing context	A nearby verb of speech, thought or perception	The frame may be distant or absent

4.4. Encoding

For exchange between projects the annotations should be expressed in a documented standard. The Text Encoding Initiative guidelines provide an element for passages spoken or thought, with attributes that state whether the passage is direct, whether it is aloud, and who its source is (TEI Consortium, 2023). Those attributes cover mode, attribution and the direct forms. The guidelines have no dedicated value for free indirect discourse, so the scheme specifies a small controlled vocabulary for form and certainty, declared in the document header and referred to through the general-purpose attributes for type and analysis. Because spans of free indirect discourse can cross sentence and paragraph boundaries, stand-off annotation that points to character offsets in a fixed transcription is recommended as the working format, with the embedded encoding generated from it. The edition and transcription used must be stated, since offsets are meaningless otherwise.

5. The Annotation Workflow

A tagging scheme is only as good as the process around it. The workflow proposed here follows the iterative model usual in language annotation, in which a draft scheme is applied to a sample, revised in the light of the problems found, and applied again before large-scale work begins (Pustejovsky and Stubbs, 2012). A pilot should use chapters from at least three novelists with different narratorial habits, so that the guidelines are not shaped by one author's manner. The guidelines themselves should be a public, versioned document containing definitions, the decision procedure, worked examples of clear and ambiguous cases, and a record of each change.

Every chapter in the main phase should be annotated independently by two people who have read the whole novel, since attribution by idiom depends on knowing how characters speak. Agreement is then measured. Two questions have to be kept apart: whether annotators marked the same stretches of text, and whether they gave matching labels to the stretches both marked. Chance-corrected coefficients such as kappa (Cohen, 1960) address the second, and Krippendorff's alpha has been extended to the first, where units are not given in advance (Krippendorff, 2004). Artstein and Poesio (2008) review the choice of coefficient and its interpretation. The verbal benchmarks proposed by Landis and Koch (1977) are

widely quoted, but they were offered as conventions and should not be treated as pass marks.

Disagreements are then adjudicated in discussion, and each is classified, following the distinctions drawn by [Gius and Jacke \(2017\)](#), as a slip, a defect in the guidelines or a real difference of interpretation. Slips are corrected, guideline defects lead to a revision, and interpretive differences are kept in the released data as ambiguous spans with both readings noted. The released resource should include the individual annotations as well as the adjudicated version, and should meet the FAIR principles of findable, accessible, interoperable and reusable data ([Wilkinson et al., 2016](#)). Because Victorian texts are out of copyright, open release of text and annotation together is straightforward.

6. Limitations and What Remains to Be Shown

This article describes a scheme and does not demonstrate it. Its central claim, that separating mode, form, attribution and certainty will yield acceptable agreement on clear cases while preserving the disputed ones, is a hypothesis. It would be tested by a pilot of the kind outlined above, reporting agreement for span boundaries and for each layer, and by a comparison with a single-label scheme applied to the same passages. It is possible that the certainty layer will itself be applied inconsistently, with one annotator's clear case being another's ambiguous one. If so, the layer would need firmer criteria, or would have to be derived from disagreement between annotators instead of being judged directly.

Other limits are built in. The four-way form distinction discards finer categories that some research questions need. The cue list is drawn from descriptions of standard literary English and may serve dialect speech poorly. The requirement that annotators read the entire novel makes the work slow and expensive, which will restrict the size of any corpus produced. And the scheme takes no side in the dispute between dual-voice and single-voice accounts, which means that it records where readers locate the form without explaining what the form is. That neutrality is deliberate, but theorists of either persuasion may find the resulting data less decisive than they would wish.

7. Conclusion

Free indirect discourse matters to the study of Victorian fiction because so much of the novels' irony and moral judgement is conducted through it, and it resists annotation for the same reason: the form works by leaving the reader to decide whose words are on the page. The scheme proposed here accepts that condition. It asks annotators for four separate judgements, gives them stated cues and an ordered procedure, records uncertainty as data, and is expressed in a standard encoding so that work done in one place can be used in another. Whether it delivers reliable and useful annotations is a question for a pilot study, which the workflow section is intended to make easy to conduct.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Artstein, R. and Poesio, M. (2008). [Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4): 555-596. <https://doi.org/10.1162/coli.07-034-R2>
- Banfield, A. (1982). [Unspeakable Sentences: Narration and Representation in the Language of Fiction](#). Routledge & Kegan Paul, Boston.
- Bray, J. (2007). [The 'dual voice' of free indirect discourse: A reading experiment](#). *Language and Literature*, 16(1): 37-52.
- Brunner, A. (2013). [Automatic recognition of speech, thought, and writing representation in German narrative texts](#). *Literary and Linguistic Computing*, 28(4): 563-575.

- Cohen, J. (1960). [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1): 37-46. <https://doi.org/10.1177/001316446002000104>
- Cohn, D. (1978). [Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction](#). Princeton University Press, Princeton, NJ.
- Fludernik, M. (1993). [The Fictions of Language and the Languages of Fiction: The Linguistic Representation of Speech and Consciousness](#). Routledge, London.
- Gius, E. and Jacke, J. (2017). [The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis](#). *International Journal of Humanities and Arts Computing*, 11(2): 233-254.
- Gunn, D.P. (2004). [Free indirect discourse and narrative authority in Emma](#). *Narrative*, 12(1): 35-54.
- Krippendorff, K. (2004). [Content Analysis: An Introduction to Its Methodology](#), 2nd edn. Sage, Thousand Oaks, CA.
- Landis, J.R. and Koch, G.G. (1977). [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1): 159-174. <https://doi.org/10.2307/2529310>
- Leech, G. and Short, M. (2007). [Style in Fiction: A Linguistic Introduction to English Fictional Prose](#), 2nd edn. Pearson Longman, Harlow.
- McHale, B. (1978). [Free indirect discourse: A survey of recent accounts](#). *PTL: A Journal for Descriptive Poetics and Theory of Literature*, 3: 249-287.
- Pascal, R. (1977). [The Dual Voice: Free Indirect Speech and its Functioning in the Nineteenth-Century European Novel](#). Manchester University Press, Manchester.
- Pustejovsky, J. and Stubbs, A. (2012). [Natural Language Annotation for Machine Learning](#). O'Reilly Media, Sebastopol, CA.
- Semino, E. and Short, M. (2004). [Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing](#). Routledge, London.
- TEI Consortium (2023). [TEI P5: Guidelines for Electronic Text Encoding and Interchange](#). Text Encoding Initiative Consortium.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al. (2016). [The FAIR Guiding Principles for scientific data management and stewardship](#). *Scientific Data*, 3: 160018. <https://doi.org/10.1038/sdata.2016.18>