Voices Speaking To and About One Another: Introducing the Project Dialogism Novel Corpus

1. Introduction

We introduce a new dataset for the computational analysis of novels: the Project Dialogism Novel Corpus (PDNC). The PDNC currently consists of 22 novels in which all quotations are identified and annotated for speaker, addressee(s), and characters mentioned. PDNC is by an order of magnitude the largest corpus of its kind. Each novel is annotated manually by a pair of annotators using customized software we developed. In addition to releasing the dataset itself alongside this paper, we are also releasing the custom annotation software we developed (including the source code) along with our annotation guidelines. In the discussion section, we present two applications of the PDNC from our own research: quote attribution and emotion dynamics. We argue that the PDNC will promote a more nuanced and accurate view of novelistic discourse; whereas much research currently envisions the novel as expressing the voice of the *author*, the PDNC presents novels as a polyphonic fabric of *characters'* voices.

2. Overview of the Project Dialogism Novel Corpus

The PDNC currently consists of 22 novels (see Table 1). In selecting 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 (REF #1) and QuoteLi (REF #15) corpora, to facilitate comparison and validation; of broad interest to a variety of scholars while still relevant to our group's interest in stylistic diversity and dialogism. 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).

The annotation workflow proceeds as follows. First, the novel is pre-processed in GutenTag (Brooke et al. 2015); from this, a provisional character list is built and likely quotations are identified. Next, the novel is manually annotated in our customized software (see Figure 1). This is done separately by two annotators. Working from our guidelines (Hammond et al. 2021), annotators select each quotation, then identify the speaker, addressee, and anyone mentioned in the quotation (whether by name or pronoun). Annotators also identify the referring expression for each quotation, as well as the quotation type: explicit (quotations in which the referring expressions give the character's name; for example, "said Emma"), pronominal (pronoun given; "she said"), or implicit (no referring expression). Once both annotators have completed their work, their annotations are compared for any discrepancies. The annotators then meet to resolve any disagreements, in what we call a "consensus exercise." Once comparison shows no disagreement between annotations, the novel is considered annotated.

The PDNC is by an order of magnitude the largest corpus of its kind (see Table 2). The largest previous corpus of novels annotated in this manner is the QuoteLi corpus, which contains only three novels (*Pride and Prejudice* and *Emma*, both in PDNC; and Chekhov's *The Steppe*, not in PDNC). The LitBank corpus includes annotations for 100 novels, but only for a very small fraction of each is annotated (on average, only 2,000 words). The Columbia Quoted Speech Attribution Corpus consists of six texts, two of which are compilations of short stories, but they are only partly annotated for quote attribution.

Table 1. *PDNC: Tokens, quotations, speakers, total # of addressees recorded, total # of mentions*

Novel	Author	# Tokens	# Quotations	# Speakers	# Addressees	# Mentions
Emma (1815)	Jane Austen	188131	2116	16	4169	412
Northanger Abbey (1817)	Jane Austen	90208	1017	16	1601	336
Persuasion (1817)	Jane Austen	96667	702	24	1678	133
Pride and Prejudice (1813)	Jane Austen	143804	1708	27	4200	2121
Sense and Sensibility (1811)	Jane Austen	139968	1545	20	3172	164
Alice's Adventures in Wonderland (1865)	Lewis Carroll	34339	1048	32	3544	84
The Man Who Was Thursday (1908)	G. K. Chesterton	69246	1357	28	4129	568
The Awakening (1899)	Kate Chopin	58925	738	18	991	981
The Mysterious Affair at Styles (1921)	Agatha Christie	72602	2226	28	7366	379
The Sign of the Four (1890)	Sir Arthur Conan Doyle	51790	891	18	1697	296
The Sport of the Gods (1902)	Paul Laurence Dunbar	50013	830	34	1370	814
The Gambler (1866; 1910)	Fyodor Dostoevsky (Trans. C. J. Hogarth)	73557	1068	20	2495	832
Howards End (1910)	E. M. Forster	136812	3131	46	5060	836
A Room with a View (1908)	E. M. Forster	83383	1989	27	3588	836
The Sun Also Rises (1926)	Ernest Hemingway	89123	3316	41	5904	2315
Daisy Miller (1879)	Henry James	26607	725	10	1209	456
Anne of Green Gables (1908)	Lucy Maud Montgomery	123465	1779	25	2412	165
A Handful of Dust (1934)	Evelyn Waugh	89070	2617	70	3628	381
The Age of Innocence (1920)	Edith Wharton	120052	1600	31	2430	714
The Invisible Man (1897)	H. G. Wells	60033	1274	29	1759	209
The Picture of Dorian Gray (1891)	Oscar Wilde	95631	1501	31	2356	1228
Night and Day (1919)	Virginia Woolf	199450	2800	38	4137	217

Figure 1. *Screen shot from our custom annotation software.*

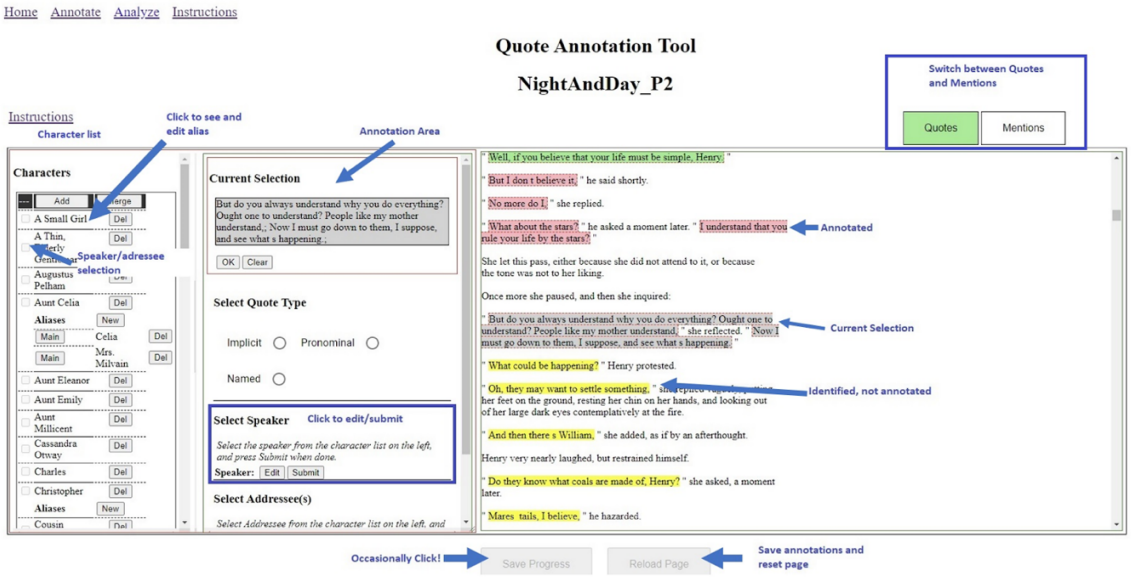


Table 2. *Comparison of PDNC with previous quotation attribution corpora*

Corpus	*Columbia Quoted Speech Attribution Corpus (2010)*	*He et al. (2013)*	*QuoteLi (2017)*	*LitBank (2020)*	*PDNC (2021)*
# Texts	6	3	3	100	22
# Annotated Quotations	3176	1901	3103	1765	35978

3. Research Applications

The research applications of the PDNC are multiple, extending well beyond the boundaries of our own research interests. Yet our own research serves to demonstrate some of its possible uses.

We began developing the PDNC primarily to test our quote attribution system (Hammond et al. 2020). The corpus has proven essential to this work, allowing us to compare our systems against state-of-the-art systems like QuoteLi and the BERT-based system in the latest release of BookNLP (see Table 3).

Table 3. *A comparison of performance of our latest quote attribution system vs. QuoteLi vs. BookNLP. Numbers reported are accuracy scores; best scores are bolded.*

		State-of-the-art			Ours	
Novel	**Author**	**# Quotations**	**Muzny et al.**	**BookNLP**	**# Quotations**	**Accuracy**
Emma (1815)	Jane Austen	1981	0.584	0.602	2102	**0.676**
Northanger Abbey (1817)	Jane Austen	1008	0.415	0.449	998	**0.702**
Persuasion (1817)	Jane Austen	631	0.569	0.605	663	**0.685**
Pride and Prejudice (1813)	Jane Austen	1696	0.563	0.561	1683	**0.682**
Sense and Sensibility (1811)	Jane Austen	1409	0.579	0.58	1529	**0.642**
Alice's Adventures in Wonderland (1865)	Lewis Carroll	951	0.894	0.92	1004	**0.959**
The Man Who Was Thursday (1908)	G. K. Chesterton	1300	0.506	0.523	1295	**0.798**
The Awakening (1899)	Kate Chopin	595	0.481	0.476	708	**0.689**
The Mysterious Affair at Styles (1921)	Agatha Christie	2121	0.162	0.209	2187	**0.60**
The Sign of the Four (1890)	Sir Arthur Conan Doyle	702	0.463	0.464	855	**0.677**
The Sport of the Gods (1902)	Paul Laurence Dunbar	766	0.403	0.373	761	**0.570**
The Gambler (1866; 1910)	Fyodor Dostoevsky	933	0.23	0.251	1038	**0.754**
Howards End (1910)	E. M. Forster	3087	0.524	0.504	3015	**0.626**
A Room with a View (1908)	E. M. Forster	1936	0.505	0.521	1954	**0.614**
The Sun Also Rises (1926)	Ernest Hemingway	2183	**0.777**	0.76	3163	0.738
Daisy Miller (1879)	Henry James	720	0.688	0.686	713	**0.833**
Anne of Green Gables (1908)	Lucy Maud Montgomery	1723	0.714	0.748	1722	**0.880**
A Handful of Dust (1934)	Evelyn Waugh	2231	0.531	**0.537**	2467	0.522
The Age of Innocence (1920)	Edith Wharton	1481	0.189	0.196	1535	**0.683**
The Invisible Man (1897)	H. G. Wells	1190	0.629	0.646	1207	**0.812**
The Picture of Dorian Gray (1891)	Oscar Wilde	1384	**0.676**	0.621	1445	0.669
Night and Day (1919)	Virginia Woolf	2783	0.58	0.631	2728	**0.689**

Perhaps the largest aim of PDNC is to reorient computational work away from conceiving novels as undifferentiated lumps of text attributed solely to their authors — but rather as complex fabrics of differentiated voices speaking to and about one another, mediated by a narrator. In the paper introducing the tool GutenTag (Hammond and Brooke 2017), one of our authors used a rudimentary version of PDNC to rebut Matthew Jockers's (2013) claim that female novelists generally write about stereotypically feminine themes. By looking at character voices *within* novels, however, rather than attributing all the novel's text to its author, we demonstrated that it was female *characters* who discussed these themes — and that Jockers's results were a secondary consequence of the fact that female authors tended to include far more female characters in their works. By allowing researchers to look *within* novels and analyze novels through the voices that make them up, PDNC will shift research away from mistaken assumptions and conclusions like Jockers's.

Our work on "emotion dynamics" — the study of change in emotional states over time — presents another example of new research enabled by the PDNC. Sentiment analysis is among the richest and most vital areas of computational literary research today. Yet major work seeking to plot novels' sentiment trajectories remains limited by the necessity of assuming a single source for all words: the author (Elsner 2012, Mohammad 2011, Jockers 2014, Reagan 2016). In a pioneering essay on "emotion dynamics" in films, Hipson and Mohammad (2021) show the benefits of considering *individual characters'* emotional trajectories. This approach enables researchers to determine each character's "home base" (typical emotional range) as well as their emotional variability and the speed at which they regulate variations. We are currently working to apply this approach to the novels in PDNC (Figures 2–4 show the emotional trajectory of Jake Barnes in Ernest Hemingway's *The Sun Also Rises*, revealing that this reputedly taciturn character in fact experiences one of the most extreme emotional troughs (in terms of valence) of any character in PDNC). We are using this approach to test whether characters' emotion dynamics track with familiar literary-critical categories such as flat vs. round characters (Forster 1927). We are also investigating the extent to which emotional trajectories are gendered, and whether male or female authors are more likely to create characters that diverge from gender norms.

Figure 2. *Emotion dynamics trajectory, valence only, for characters in Ernest Hemingway's The Sun Also Rises. Jake Barnes's emotional trajectory is highlighted; the trough three-quarters of the way through the novel (~ 76%-87%) occurs during and after his fight with Robert Cohn at the Fiesta.*

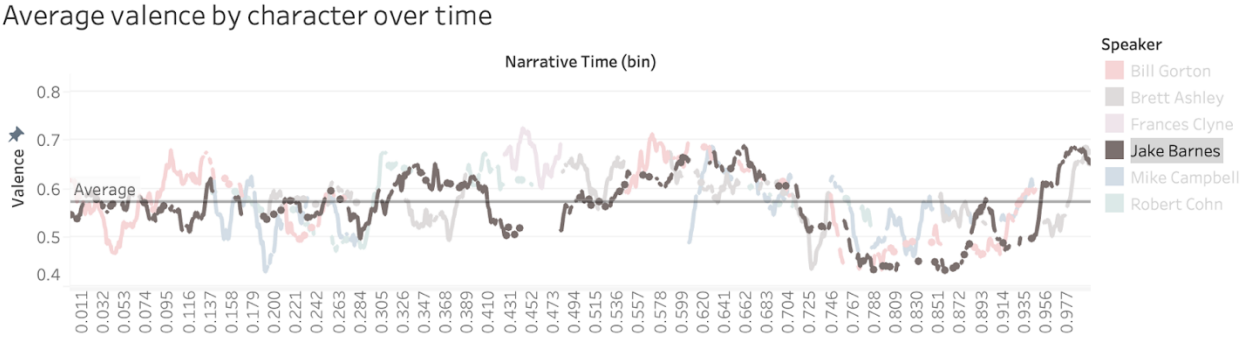


Figure 3. *Emotion dynamics, valence only, for all characters in PDNC. Jakes Barnes's trajectory (highlighted) is extreme in the context of the characters in our corpus.*

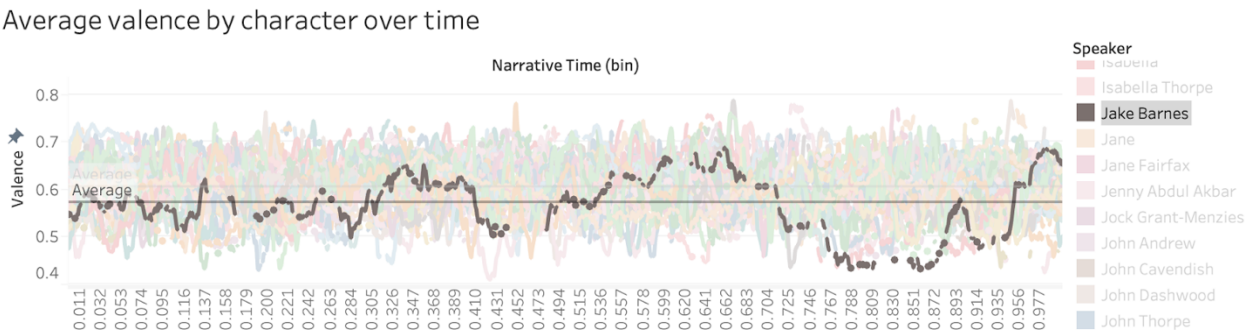
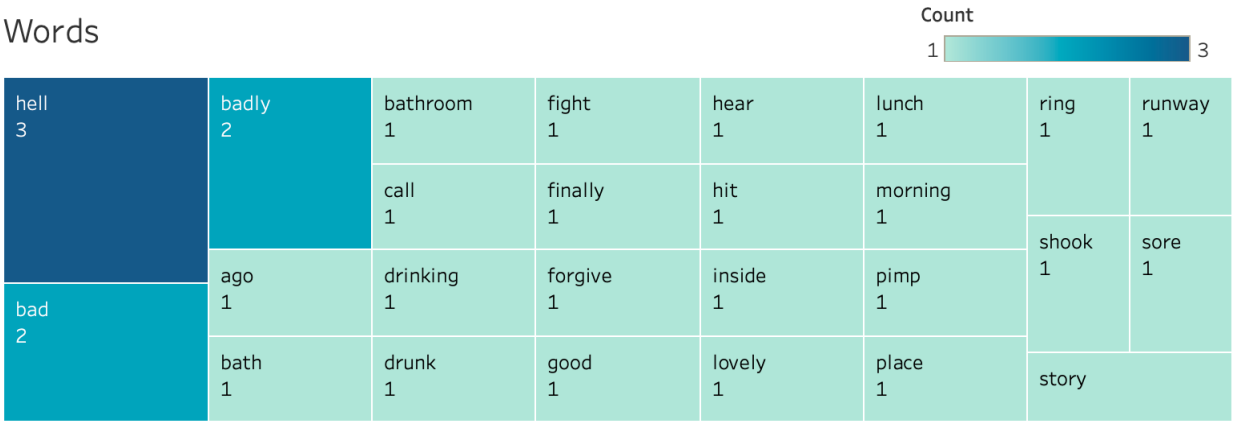


Figure 4. *Emotion words (with frequency count) used by Jake Barnes during trough (76%-87% portion of novel)*



Works Cited

1. Bamman, David, Sejal Popat, and Sheng Shen. "An annotated dataset of literary entities." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2138-2144. 2019.
2. Brooke, Julian, Adam Hammond, and Graeme Hirst. "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*, pp. 42-47. 2015.
3. Elsner, Micha. "Character-based kernels for novelistic plot structure." In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 634-644. 2012.
4. Elson, David K., and Kathleen R. McKeown. "Automatic attribution of quoted speech in literary narrative." In *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.
5. Forster, E. M. *Aspects of the Novel.* 1927.

6. Hammond, Adam, and Julian Brooke. “GutenTag: A User-Friendly, Open-Access, Open-Source System for Reproducible Large-Scale Computational Literary Analysis.” In *Proceedings of the Digital Humanities 2017 Conference*, pp. 246–249. 2017.
7. Hammond, Adam, Krishnapriya Vishnubhotla, and Graeme Hirst. “The Words Themselves: A Content-Based Approach to Quote Attribution.” *Proceedings of the Digital Humanities 2020 Conference.* 2020.
8. Hammond, Adam, Krishnapriya Vishnubhotla, Leah Duarte, Sanghoon Oh, Jovana Pajovic, and Beck Siegal. “Annotation Guidelines for the Project Dialogism Novel Corpus.” 2021. https://tinyurl.com/quoteattribution
9. Mohammad, Saif. "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174-184. 2018.
10. Brooke, Julian, and Graeme Hirst. "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*, pp. 673-679. 2013.
11. He, Hua, Denilson Barbosa, and Grzegorz Kondrak. "Identification of speakers in novels." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1312-1320. 2013.
12. Hipson, Will E, and Saif M Mohammad. “Emotion dynamics in movie dialogues.” *PloS one* vol. 16,9 e0256153. 20 Sep. 2021, doi:10.1371/journal.pone.0256153
13. Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History* (University of Illinois Press). 2013.
14. Jockers, Matthew. “A novel method for detecting plot.” 2014. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/
15. Mohammad, Saif. “From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales.” In *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH). 2011.
16. Muzny, Felix, Michael Fang, Angel Chang, and Dan Jurafsky. "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*, pp. 460-470. 2017.
17. Reagan, Andrew J, Lewis Mitchell, Dilan Kiley, Christopher M Danforth & Peter Sheridan Dodds. “The emotional arcs of stories are dominated by six basic shapes.” *EPJ Data Science* 5(31), pp. 1–12. 2016.