

Lattice (CNRS – École Normale Supérieure – Université Sorbonne Nouvelle), Paris, France
antoine.bourgois@ens.psl.eu

Diagram illustrating the interconnected nature of three fields:

- Computational Literary Studies
- Cultural Analytics
- Digital Humanities

Downstream Processes

[1] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. *OntoNotes* release 5.0

[2] Andreas van Cranenburgh. 2019. A dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

[3] David Bamman, Olivia Lewke, and Anya Mansor. 2020. An annotated dataset of coreference in English Literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association

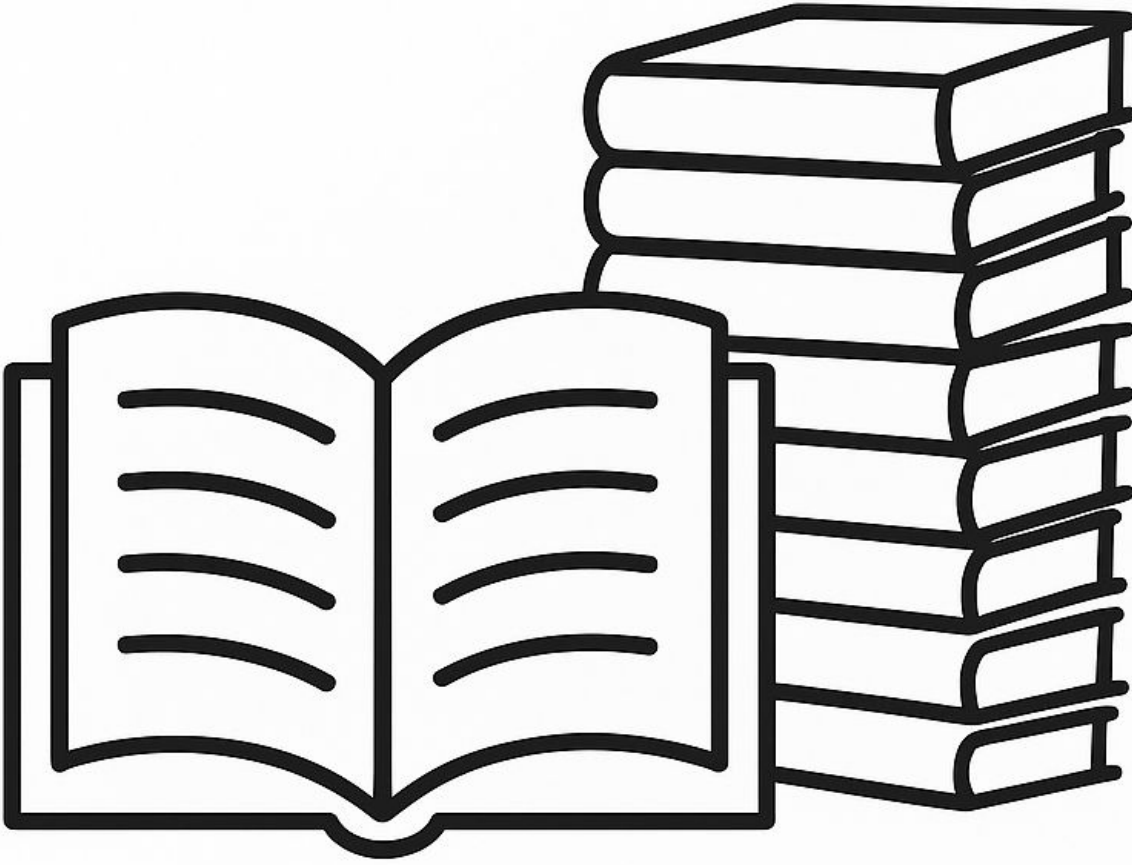
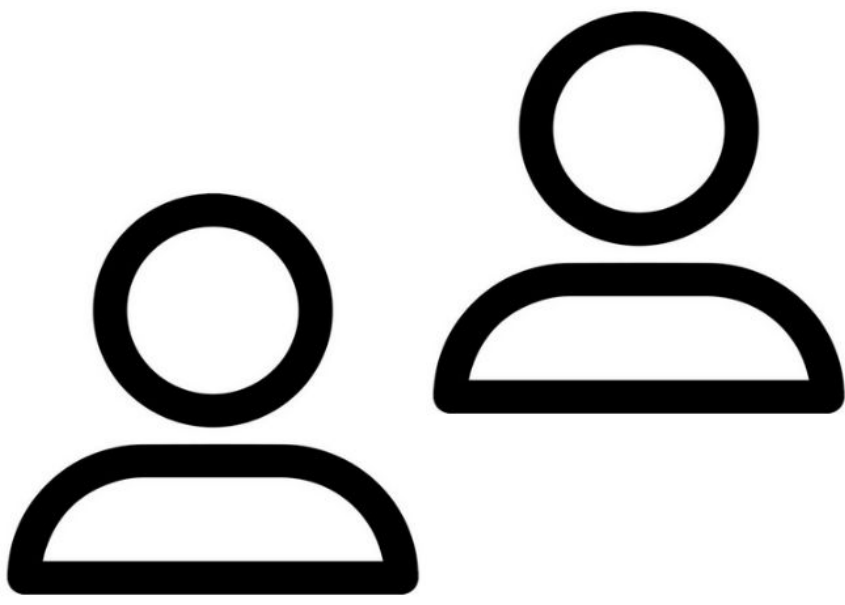
[4] Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Ko-conovel: Annotated dataset of character coreference in korean novels.

[5] Frédéric Mélanie, Jean Barré, Olga Semnink, Clé-mont Plançon, Marco Naguib, Marcial Pastor, and Thierry Poibeau. 2024. Booknp-fr, the french versant of booknp: a tailored pipeline for 19th and 20th century french literature. *Journal of Computational Literary Studies*.

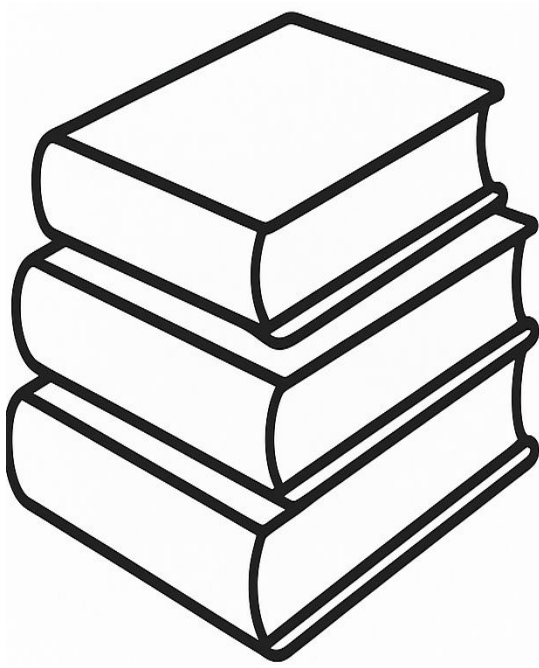
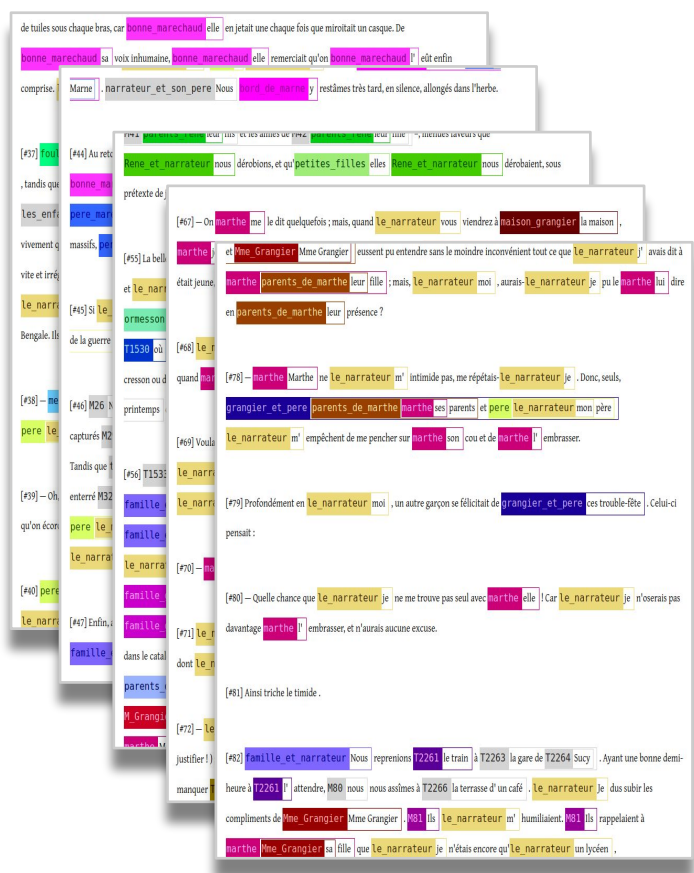
[6] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. *Association for Computational Linguistics*.

[7] Giuliano Martinielli, Tommaso Bonomo, Pere-Lluís Huguet Cabot, and Roberto Navigli. 2025. BOOKCOREF: Coreference Resolution at Book Scale. Vienna, Austria. Association for Computational Linguistics.

Dataset	Lang.	Domain	Doc.	Tokens	Tokens / Doc.	
					Average	Maximum
Annotated Datasets						
OntoNotes (en) – Weischedel et al., 2013	English	Non-literary	3,493	1,600,000	467	4,009
DROC – Krug et al., 2018	German	Fiction	90	393,164	4,368	15,718
RiddleCoref – Cranenburgh et al., 2019	Dutch	Fiction	21	107,143	5,102	–
LitBank – Bamman et al., 2020	English	Fiction	100	210,532	2,105	3,419
FantasyCoref – Han et al., 2021	English	Fantasy	214	367,891	1,719	13,471
KoCoNovel – Kim et al., 2024	Korean	Fiction	50	178,000	3,578	19,875
LitBank-fr – Mélanie et al., 2024	French	Fiction	28	275,360	9,834	30,987
Target Datasets						
Standard Ebooks	English	Fiction	770	82,855,210	107,604	1,105,964
Chapitres – Leblond et al., 2022	French	Fiction	2,960	240,971,614	81,409	878,645
Contribution (Ours)						
Ours	French	Fiction	3	285,176	95,058	115,415



Dataset
French National Library - Fictions
(BNF)GALLICA_FILTERED_output_2025_08
files_count: 12,011
overall_tokens: 1,188,297,735



Corpus (test set)	Model	Mentions	Tokens / Doc	MUC	B ³	CEAF _e	CoNLL
LitBank-fr (test-set)	End-to-end Mélanie (2024)	Gold	2,000	88.0	69.2	71.8	76.4
LitBank-fr (test-set)	Ours	Gold	2,000	92.43	70.67	75.59	79.56
Long-LitBank-fr (3 docs)	Ours	Predicted	93,019	95.59	45.40	35.95	58.98
Cross-datasets							
LitBank (English)	Bamman (2020)	Predicted	2,105	84.3	62.73	57.3	68.1
LitBank (English)	Thirukovalluru (2021)	Predicted	2,105	89.50	78.21	67.59	78.44
LitBank-fr (LOOCV)	Ours	Predicted	2,105	84.58	74.77	63.30	73.21
BookCoref (gold)	Longdoc	Predicted	76,419	93.5	62.4	45.3	67.0
BookCoref (gold)	Maverick-xl	Predicted	76,419	94.3	55.3	33.4	61.0
Long-LitBank-fr (3 docs)	Ours	Predicted	76,000	94.99	47.51	37.49	60.00

Corpus	Model	Mentions	Tokens / Doc	MUC	B ³	CEAF _e	CoNLL
LitBank (English)	Bamman (2020)	Predicted	2,105	84.3	62.73	57.3	68.1
LitBank (English)	Thirukovalluru (2021)	Predicted	2,105	89.50	78.21	67.59	78.44
LitBank-fr (LOOCV)	Ours	Predicted	2,105	84.58	74.77	63.30	73.21
BookCoref (gold)	Longdoc	Predicted	76,419	93.5	62.4	45.3	67.0
BookCoref (gold)	Maverick-xl	Predicted	76,419	94.3	55.3	33.4	61.0
Long-LitBank-fr (LOOCV)	Ours	Predicted	76,000	94.99	47.51	37.49	60.00

