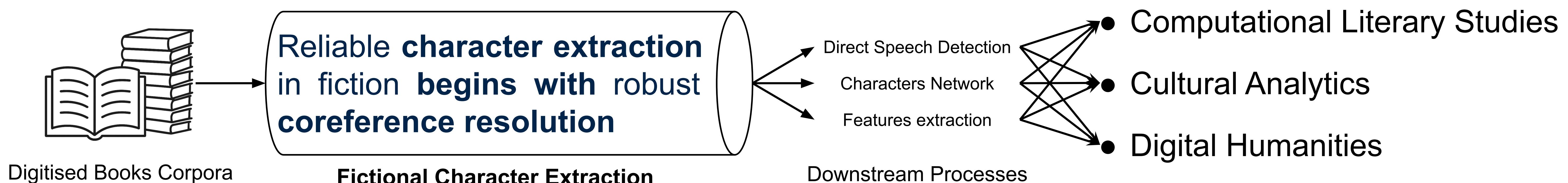


The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works

Antoine Bourgois and Thierry Poibeau

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

Context: Coreference Resolution for Computational Narratology

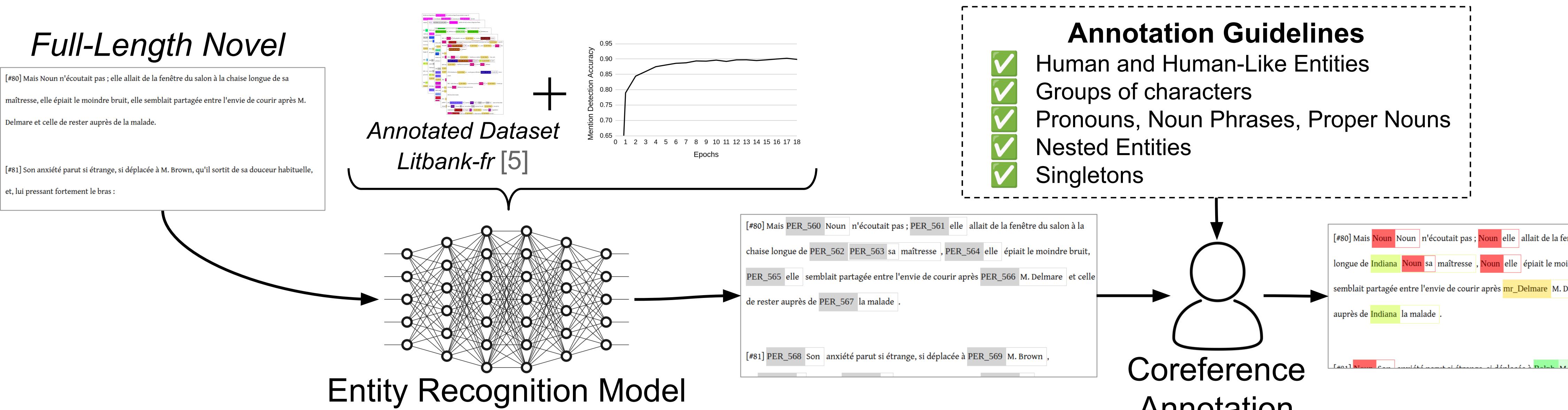


Data Mismatch: Annotated Dataset VS Downstream Applications

Datasets	Lang.	Domain	Doc.	Tokens / Doc.	
				Average	Maximum
Annotated Datasets					
OntoNotes ^{en} – 2013 [1]	English	Non-literary	3,493	467	4,009
RiddleCoref – 2019 [2]	Dutch	Fiction	21	5,102	–
LitBank – 2020 [3]	English	Fiction	100	2,105	3,419
KoCoNovel – 2024 [4]	Korean	Fiction	50	3,578	19,875
LitBank-fr – 2024 [5]	French	Fiction	28	9,834	30,987
Target Datasets					
Project Gutenberg	English	Fiction	61,300	110,000	1,105,964
National Library of France - Gallica	French	Fiction	15,000	100,000	878,645

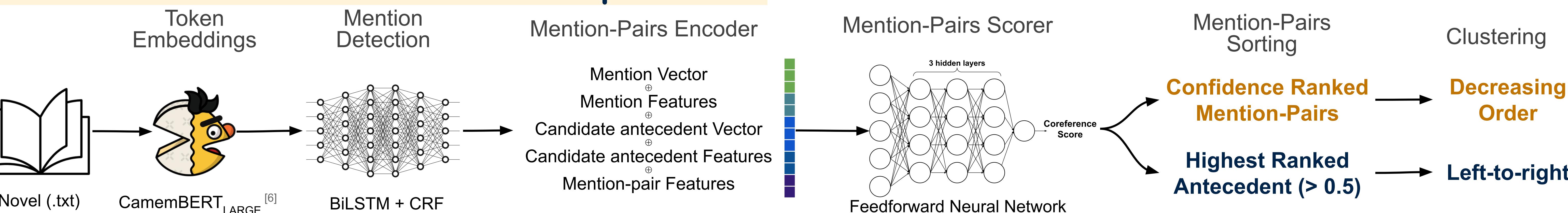
Existing annotated fiction coreference datasets are, on average, **10x shorter** than the narratives used in real-world applications.

New Annotated Dataset

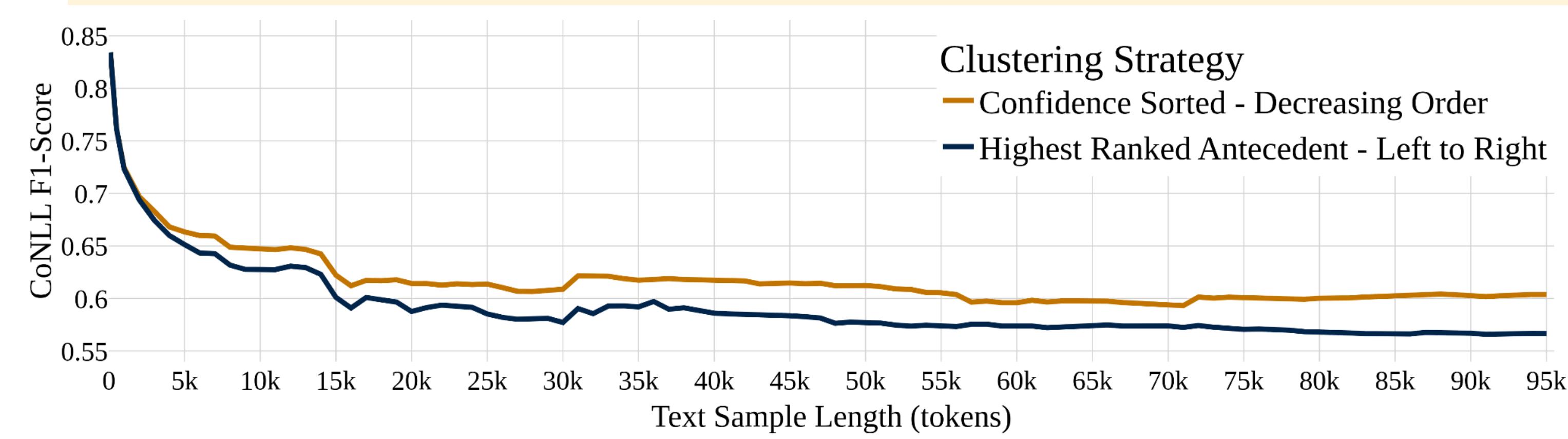


Documents	3
Average Tokens / Doc.	95,058
Average Mentions / Doc.	13,178
Coreference Chains / Doc.	159
Average Mentions / Chain	82
Maximum Mentions / Chain	4,932
Average Entity Spread (tokens)	17,529
Maximum Entity Spread (tokens)	115,369
Plural Mentions Ratio	8.13 %
Proper Mentions	12.79 %
Nominal Mentions	12.26 %
Pronominal Mentions	74.95 %

Coreference Resolution Pipeline



Document Length Impact on Coreference Resolution Performance



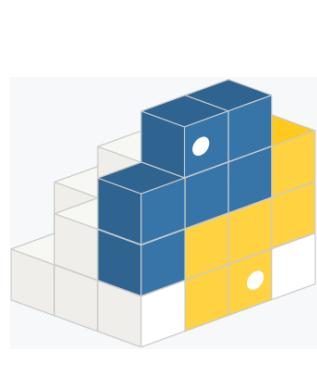
Corpus (test set)	Model	Mentions	Tokens / Doc	MUC	B ³	CEAFe	CoNLL
LitBank-fr (test-set)	End-to-end [x]	Gold	2,000	88.0	69.2	71.8	76.4
LitBank-fr (test-set)	Ours	Gold	2,000	92.43	70.67	75.59	79.56
Long-LitBank-fr (3 docs)	Ours	Predicted	93,019	95.59	45.40	35.95	58.98
Cross-datasets							
BookCoref _{gold} [7]	Longdoc [x]	Predicted	76,419	93.5	62.4	45.3	67.0
BookCoref _{gold} [7]	Maverick-xl [x]	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

Perspectives

- Multilingual adaptation: English, Italian, Russian
- Downstream applications: Character archetypes, Gender Study

Available Now

```
pip install propp_fr
import propp_fr
propp_fr.process_text_file("your_novel.txt")
```



Character Mention Detection Model
AntoineBourgois/propp-fr_NER_camembert-large_PER
Mention-Pairs Scorer
.../propp-fr_coreference-resolution_camembert-large_PER