OpenBoek: A Corpus of Coreference and Entities in Dutch Literature

Frank van den Berg, Esther Ploeger, Menno Robben,
Pauline Schomaker, Robin Snoek, Remi Thüss,
Andreas van Cranenburgh a.w.van.cranenburgh@rug.nl
CLCG, University of Groningen

The literary domain presents unique challenges for NLP tasks such as coreference resolution. Addressing these challenges requires annotated data of sufficient quantity and quality for training and evaluating models. Recent work introduced coreference datasets for classic English novels (Litbank; Bamman et al., 2020) and contemporary Dutch novels (RiddleCoref; van Cranenburgh, 2019).[1] Unfortunately, RiddleCoref is encumbered by copyright; i.e., the annotated texts cannot be made available.

We address this by annotating a corpus of Dutch public domain novels which we will release under an open license; see Table 1 and 2 for an overview of the corpus. The OpenBoek corpus currently consists of 9 fragments of Project Gutenberg texts, both translated and original Dutch novels. We annotated the full text of the novellas by Nescio, while the other fragments are initial segments of 10k tokens. The documents are therefore considerably longer than those in typical coreference corpora (SoNaR-1: 1k, Litbank: 2k). This fragment length was chosen with the aim of evaluating and tackling the particular challenges of long-document coreference resolution.

Annotation proceeded with the same semi-automatic method and annotation guidelines[2] as RiddleCoref: the texts were parsed by Alpino and coreference output of dutchcoref (van Cranenburgh, 2019) was manually corrected by multiple annotators. Mentions are manually corrected and exclude non-referring expressions. See example (1).

In addition to coreference, we annotated the gender (neuter, female, male, gendered but mixed/unknown) and number (singular, plural) of all entities in RiddleCoref and OpenBoek. The gender attribute also distinguishes person and non-person entities. Although the syntactic gender, number available in Alpino parse trees is already informative, we annotated the semantic gender and number; e.g., the grammatically neuter *het meisje* is annotated as female and the singular *de groep* is annotated as plural due to being a collective noun. These annotations are useful for training models to detect mention features for coreference resolution, among other possible applications.

Keywords: coreference, literature, annotation, dataset, Dutch

[1] The 1M word SoNaR-1 corpus (Schuurman et al., 2010) contains 2000 tokens of coreference annotations for books, but the majority consists of Wikipedia and various other genres.

[2] https://github.com/andreasvc/dutchcoref/blob/master/annotationguidelines.pdf

(1) Toen had [zij]$_1$ [Henri Van Raat]$_2$ ontmoet, en sedert verbaasde [zij]$_1$ [zich]$_1$ vaak, hoe [die goede lobbes]$_2$, zooals [zij]$_1$ [hem]$_2$ noemde, [die]$_2$ toch zoo weinig op [den held [[harer]$_1$ droomen]$_3$]$_4$ geleek, [zooveel sympathie]$_5$ in [haar]$_1$ verwekte, dat [zij]$_1$ dikwijls, plotseling, naar [[zijn]$_2$ bijzijn]$_6$ verlangen kon. (Couperus, *Eline Vere*)

documents	9	mentions / entities	2.66
sentences	5,709	mentions / tokens	0.228
sents per doc	634.3	entities / tokens	0.0857
avg sent len	18.1	% pronouns	40.9
mentions	23,650	% nominals	48.0
entities	8,875	% names	11.1

Table 1: Corpus statistics.

Author, title	# tokens	# n	f	m	fm	sg	pl
Conan Doyle, De Agra Schat	10,536	792	7	34	60	712	181
Couperus, Eline Vere	10,473	746	42	32	55	638	237
Hugo, De Ellendigen	10,488	807	13	65	178	731	332
Multatuli, Max Havelaar	10,646	627	27	34	112	611	189
Nescio, De Uitvreter	15,210	1067	10	48	141	955	311
Nescio, Dichtertje	18,245	1296	55	48	151	1150	400
Nescio, Titaantjes	12,538	689	16	38	68	569	242
Tolstoy, Anna Karenina	10,579	606	21	34	74	553	182
Verne, ReisOmDeWereld	10,516	738	4	33	157	680	252
Total	103,522	7368	195	366	996	6599	2326

Table 2: Corpus composition (# tokens, # entities).

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of LREC*, pages 44–54. `https://aclweb.org/anthology/2020.lrec-1.6`.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In *Proceedings of LREC*, pages 2471–2477. `https://aclweb.org/anthology/L10-1104`.

Andreas van Cranenburgh. 2019. A Dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54. `https://clinjournal.org/clinj/article/view/91`.