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

university of groningen

A Corpus of . . .

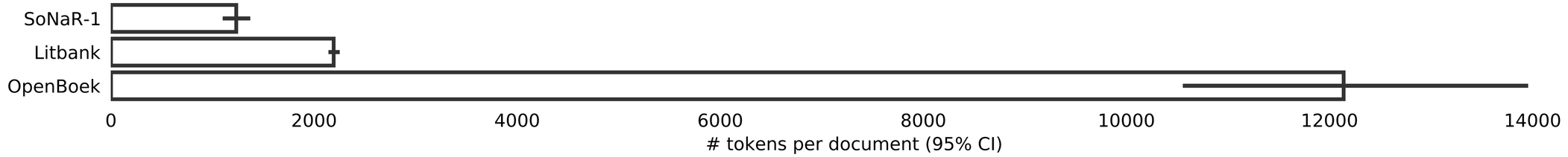
- Coreference & entity features (108k tokens, 23k mentions)
- Classic literature (Public Domain)
- Long documents (> 10k tokens)

Coref Example: Nescio, *Dichtertje*

Tweemaal schudde de God van Nederland zijn eerbiedwaardig hoofd en tweemaal schoven z'n eerbiedwaardige grauwe bakkebaarden heen en weer over z'n vest.

Entity features

gender	number	mentions
fm	sg	Men, zich, men
n	sg	de , tot kleedkamer ingerichte , eetzaal, de kamer, der kamer
n	sg	een psyché
f	sg	Frédérique Van Erlevoort, ze, haar, haar, . . .



Annotation Procedure

- Alpino: parse trees
- dutchcoref: coreference output
- CorefAnnotator: correction of mentions and coref clusters
- Spreadsheet: entity features (female, male, mixed, neuter; singular, plural)

https://andreasvc.github.io/openboek

Corpus composition & stats

Author, title	# tokens	% n	f	m	fm	sg	pl
Conan Doyle, De Agra Schat	10,536	89	1	4	7	80	20
Couperus, Eline Vere	10,473	85	5	4	6	73	27
Hugo, De Ellendigen	10,488	76	1	6	17	69	31
Multatuli, Max Havelaar	10,646	78	3	4	14	76	24
Nescio, De Uitvreter	15,210	84	1	4	11	75	25
Nescio, Dichtertje	18,245	84	4	3	10	74	26
Nescio, Titaantjes	12,538	85	2	5	8	70	30
Tolstoy, Anna Karenina	10,579	82	3	5	10	75	25
Verne, ReisOmDeWereld	10,516	79	0	4	17	73	27