AIUCD 2021

DNT: a Diachronic and Multi-genre Corpus of English Texts

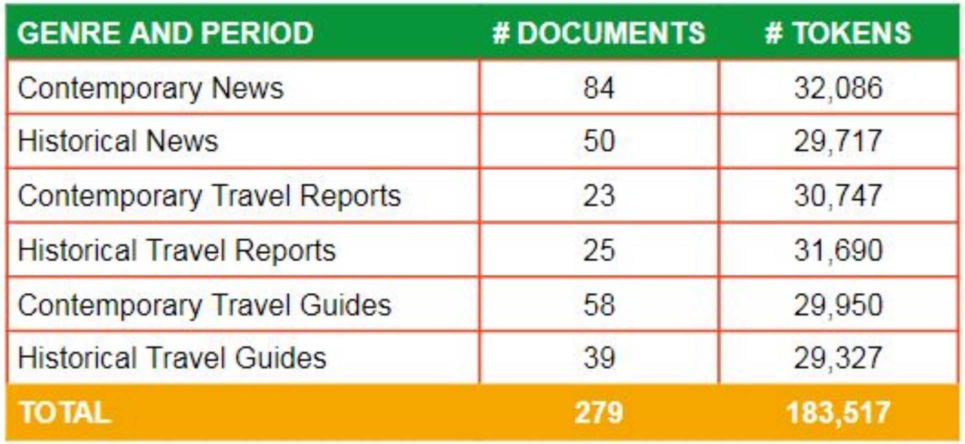
rijksuniversiteit groningen

T. Caselli - R. Sprugnoli

UNIVERSITÀ CATTOLICA del Sacro Cuore

DNT = Diachronic News and Travel corpus

- 3 genres: news, travel reports, travel guides
- 2 temporal periods: 1860-1939 & 1998-2017
- clean texts: no OCR errors
- multi-layer annotations
- https://github.com/tommasoc80/DNT



GENRE AND PERIOD	# DOCUMENTS	# TOKENS
Contemporary News	84	32,086
Historical News	50	29,717
Contemporary Travel Reports	23	30,747
Historical Travel Reports	25	31,690
Contemporary Travel Guides	58	29,950
Historical Travel Guides	39	29,327
TOTAL	279	183,517

USE CASES

- Linguistic annotations, NLP applications
- All annotations and models are available online

CONTENT TYPES CLASSIFICATION

- Manual annotation and automatic classification of micro illocutionary acts at clause level
- New annotation scheme
- Cross-genre and cross-time analysis

EVENT PROCESSING

- Manual annotation and automatic identification + classification of events in historical news, travel reports and guides
- New annotation scheme
- Cross-genre analysis

PLACE NAMES RECOGNITION

- Manual annotation and automatic identification of geographical (*Vesuvius, Mediterrean Sea*), political (*Tuscany, Regno delle due Sicilie*) and functional (*Church of St. Severo, Forum Romanum*) locations in travel reports and guides

GENRE AND PERIOD	# DOCUMENTS	# TOKENS
Contemporary News	84	32,086
Historical News	50	29,717
Contemporary Travel Reports	23	30,747
Historical Travel Reports	25	31,690
Contemporary Travel Guides	58	29,950
Historical Travel Guides	39	29,327
TOTAL	**279**	**183,517**