APPENDIX A Template Data Report Version July 2020

Title	CACAPO dataset

Organization

Author	van der Lee, Chris[1] Emmery, Chris[2] Wubben, Sander[1] Krahmer, Emiel[1]
Affiliation	Tilburg University, Tilburg School of Humanities and Digital Sciences, Department of Communication and Cognition[1] Tilburg University, Tilburg School of Humanities and Digital Sciences, Department of Cognitive Science and Artificial Intelligence[2]
Email	C.vdrLee@tilburguniversity.edu
Distributor	DataverseNL
Ethical clearance	N/A
Preregistration	N/A

Research context

Description	The **C**ombinations of **A**ligned data-senten **C**es from n **A**turally **Pr** **O**duced texts (hereafter: CACAPO) dataset is a dataset for data-to-text generation. The dataset contains over 20,000 sentences from automatically scraped news reports for the sports, weather, stock, and incidents domain in English and Dutch, aligned with relevant attribute-value paired data. To our knowledge, this is the first dataset based on "naturally occurring" human-written texts (i.e., texts that were not collected in a task-based setting), that covers various domains, as well as multiple languages.
Kind of data	Annotated corpora
Publication	van der Lee, C., Emmery, C., Wubben, S., & Krahmer, E. (2020, December). The CACAPO dataset: A multilingual, multi-domain dataset for neural pipeline and end-to-end data-to-text generation. In Proceedings of the 13th International Conference on Natural Language Generation (pp. 68-79). url: https://aclanthology.org/2020.inlg-1.10
Keywords	Tilburg University, Humanities, Computation and Language, Natural Language Generation, Data-to-Text Generation

Data production

Producer	Tilburg University
Production date	Begin date: 22-08-2017 – End date: 27-02-2020
Method	The texts were collected using automatic scrapers or an interface that allowed quick collection of the article. Aligned data was manually annotated by two annotators.
Universe	News reports on traffic/gun violence incidents, soccer/baseball matches, stocks, and weather, published between 2016 and 2019.
Data sources	A variety of news websites. A full list of articles and sources can be found on https://github.com/TallChris91/CACAPO-Dataset .

Country / Nation	The reports come from Dutch- and English-speaking countries. Mostly The Netherlands, United Kingdom, and United States of America.

Rights and restrictions

Restrictions (file permission)	Public
License	CCBY
Rights	N/A
Data retention period	10 years.

Read-me

Data files	1. Full_Dict_NL.json, Full_Dict_EN.json, Phrase_Dict.json, PhraseTable.json → JSON files that contain information on verbs and determiners. This is useful for a realization module to apply the correct word form in a given situation. 2. WebNLGFormatTrain.xml, WebNLGFormatDev.xml, WebNLGFormatTest.xml → The corpus files in XML format. Their structure is the same as (enriched) WebNLG's structure v 1.4 (see https://github.com/ThiagoCF05/webnlg).
Supplemental material	N/A
Structure data package	1. nl a. Incidents i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. Phrase_Dict.json b. Sports i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. Phrase_Dict.json c. Stocks i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. Phrase_Dict.json d. Weather i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. Phrase_Dict.json e. Full_Dict_NL.json 2. en a. Incidents i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. PhraseTable.json b. Sports i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml

	iii. WebNLGFormatTest.xml iv. PhraseTable.json c. Stocks i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. PhraseTable.json d. Weather i. WebNLGFormatTrain.xml ii. WebNLGFormatDev.xml iii. WebNLGFormatTest.xml iv. PhraseTable.json e. Full_Dict_EN.json