Published in CommonLit

Michelle Brown Follow

Dec 6, 2021 · 4 min read · Listen

Introducing: The CLEAR Corpus, an open dataset to advance research

CommonLit collaborated with Georgia State University to create an dataset of almost 5,000 reading passages to support a wide range of research.

About the CLEAR Corpus

The CLEAR (CommonLit Ease of Readability) Corpus is an open dataset of almost 5,000 reading passage excerpts which have been curated for research purposes. The passages were selected with a grade 3–12 English Language Arts classroom context in mind, but we hope that they will prove useful in a wide range of research fields.

The following Readability indices are provided for each excerpt: Flesch-Reading-Ease, Flesch-Kincaid-Grade-Level, Automated Readability Index, SMOG Readability, New Dale-Chall Readability Formula, CAREC, CAREC_M, CARES, CML2RI, and First Place through Sixth Place Predictions from the Kaggle Readability Prize.

An early publication describing the initial corpus and its development can be found here.

On CommonLit.org, students interact with digital text. The CommonLit Library includes over 2,500 free lessons for grades 3–12.

Development of the Corpus

Development of this dataset was a joint project between CommonLit and the Georgia State University department of Applied Linguistics & ESL. The project was made possible through the support of Schmidt Futures.

The CLEAR Corpus was curated by CommonLit's team, meticulously reviewed and tagged with metadata, and first used in the Kaggle Readability Prize competition in 2021. Through the Kaggle competition, thousands of teams around the world spent months developing readability complexity measurement algorithms. The winning algorithms of the 2021 competition were openly released and the output has been included in the dataset.

- **ID —** A unique identifier for each Excerpt.
- **Last Changed —** Optional field. The corpus version in which any data in this row last changed. Present if the row changed since v6.0.
- **Author —** The author of the Excerpt.
- **Title —** Copied directly from the data source cited in URL.
- **Anthology —** Optional field. Present if source appeared within a collection.
- **URL —** The URL from which the Excerpt was copied.
- **Source —** Derived from URL.
- **Pub Year —** Copyright year of the Excerpt.
- **Category —** Informational or Literary.
- **Location —** Start if Excerpt begins the passage. End if the Excerpt concludes the passage. Otherwise Mid.
- **License —** Restrictions for use or distribution of the Excerpt. See below.
- **MPAA Max —** The highest of the MPAA ratings from two trained raters.
- **MPAA #Max —** MPAA Max expressed as an integer.
- **MPAA #Avg —** The MPAA average of the two raters.
- **Excerpt —** The excerpt which was read and rated by the MPAA raters, teacher raters, traditional indices, and Kaggle competitors.
- **Google WC —** Word count formula in Google Sheets =COUNTA(SPLIT(Excerpt," "))
- **Joon WC v1 —** An alternative word counter.
- **British WC —** A count of uniquely British words (non-American spelling).



- **Sentence Count v2 —** Sentence count from algorithm 2. The dual sentence counters vary in their counting of multiple sentences within a single quotation, and other such unusual situations. They are both provided as estimates.
- **Paragraphs —** Paragraph count.
- **BT Easiness —** Bradley-Terry coefficient derived from teacher ratings from teacher raters. See EDM21 Section 4.5.
- **BT s.e. —** Standard error associated with BT_easiness.
- **Readability indices —** Nine traditional indices which were generated by ARTE v1.1: Flesch-Reading-Ease, Flesch-Kincaid-Grade-Level, Automated Readability Index, SMOG Readability, New Dale-Chall Readability Formula, CAREC, CAREC_M, CARES, CML2RI.
- **Readability predictions —** Six additional predictions which resulted from the Kaggle competition.
- **Kaggle split —** The Kaggle Readability Prize data set containing the Excerpt. The competition used a 60–40% train-test split.

Licensing and Use Restrictions

This dataset and metadata developed by our team are distributed under an open MIT license, and we encourage you to make use of it in your research. However, please take note of the license under which each passage excerpt is distributed before republishing the corpus.

Most CLEAR passages are in the public domain, and therefore carry no restrictions. Many licensed passages are distributed under either CC BY 4.0 or CC BY-SA 3.0. Other Creative Commons licenses are described on their site. At a high level, “BY” licenses require that when an excerpt is reused, attribution must be included to the original. “NC” denotes a non-commercial restriction, limiting the republication of the excerpt to only non-commercial purposes. Finally, “SA” requires that reuse or derivative works including the



CommonLit will maintain and continuously improve the CLEAR Corpus. Please use this form to submit any feedback you would like our team to consider. New modifications to the corpus will be incorporated once or twice per year. The spreadsheet in the first paragraph of this announcement will always display the most current version of the corpus and the version number will be reflected in the name of the file.

Release History

6.0 — May 2021 — Essential columns (only) were released during the Kaggle competition.

6.01 — December 2021 — The first general release of the CLEAR corpus. Included 138 row-level corrections, mostly to Publication Year and License.

Contact

We hope that this dataset can foster new research and innovation in the field of education, natural language processing, and beyond. If you have any questions, or would like to share your research incorporating the CLEAR Corpus with us, please contact info@commonlit.org.