L2WS 2021

L2WS (L2 Written Summary) 2021 is a corpus containing 1 source text and 40 student summaries manually segmented in Idea Units.

Description

The source text is an expository text with problem solution structures embedded in it. The summaries were written by 40 undergraduate students at a university in Japan. They were all non-native speakers of English. The summaries were collected as part of an assignment for an academic writing course in which the students were asked to read the source text (391 words) and summarise its main ideas and key details in approximately 80 words.

Citation

Please cite the following paper when using the L2WS 2021 corpus in your study:

- Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki and Mika Ishizuka. *Automating Idea Unit Segmentation and Alignment for Assessing Reading Comprehension via Summary Protocol Analysis*. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pp. 4663-4673. Marseille, France, June 2022.

File Structure

In the repository, there are two folders. One for the source text and one for the summaries.
`source_text/raw.txt` is the plaintext version of the source text.
`source_text/gold.txt` is the manual annotation of the source text.
The source text was annotated by the two annotators in a joint session, producing a single gold standard.
The 40 summaries were annotated by both annotators individually and later merged into a double-coded gold standard via consensus.
The plaintext of the 40 summaries is found in `summaries/raw`.
`summaries/annotator1` and `summaries/annotator2` are the directories for single-coded annotations.
`summaries/gold` contains the double-coded gold standard.

Syntax

Each line of an annotated document represents an Idea Unit.
Discontinuous Idea Units are prefixed by an index and the special character `|`.
Lines beginning with the same index represent one discontinuous Idea Unit.