ICLE_1300 metadata documentation

The ICLE_1300.csv and ICLE_1300.xlsx files provide basic text metadata and comparative judgement (CJ) scores for 1300 argumentative texts from the International Corpus of Learner English (ICLE). The texts themselves (along with extended metadata) are available via the ICLE website (https://corpora.uclouvain.be/cecl/icle/).

This document describes the sampling of the 1300 texts from the corpus, and the process of generating CJ scores for them.

Text pre-processing

The ICLE contains texts written in response to several hundred different essay prompts. In some cases, there are prompts which differ only in minor ways, such as their wording, spelling, or punctuation. In order to maximise comparability between similar essays, we created "theme" categories which collected essays on the same topic. There are "macro" themes, which comprise essays on the same general theme but have distinct foci; and "micro" themes, which comprise essays which represent differing wordings of essentially the same prompt.

Text selection

The texts were selected in three stages. The first stage involved the selection of 500 texts which would be used both in CJ tasks within the framework of the CLAP project, and a parallel rubric-based assessment task (Thwaites et al., 2024).

In this first stage, three principles for text selection were identified, with the general goal of selecting texts that are representative of the ICLE argumentative corpus as a whole:

1. Text length
    - The mean length of texts in the corpus was 598 words. Only texts which were one standard deviation lower or higher than this number were selected.
2. Essay prompt
    - Some of the prompts included in the ICLE occur much more frequently than others. We aimed to select from among the more frequent prompts, since this allows researchers to compare texts which have the same prompt but different proficiency and/or L1 background. Therefore, only texts whose prompts occurred at least 50 times in the corpus, across at least 3 L1 groups, were selected.
3. L1 background
    a. The ICLE corpus contains texts collected from English learners in 25 countries. We aimed to select 20 texts from each of these L1 backgrounds.
    b. Additionally, for each L1 group, we aimed for these 20 texts to consist of 4 texts each from 5 of the previously selected essay prompts.

It was possible to select 500 texts according to principles 1-3a. However, due to variance in the data, some flexibility was required on principle 3b. For example, for some L2 groups very few different prompts were used, meaning that that group's 20 texts consist of responses to only 1 or 2 prompts. On the other hand, for other groups very few prompts were used frequently, meaning that the 20 texts contained up to seven different prompts.

After selecting the initial 500 texts, the second stage was to ensure that an earlier sample of 222 texts was included in the current dataset. These were texts that had, during a previous study, been manually annotated to tag language errors. The selection criteria used in that study were different to those outlined above; the texts were all written by L1 Spanish, French, and German speakers, and covered a range of topics (including some not present in the above sample of 500 texts). Since only 17 of these texts were included in the above sample of 500 texts, the remaining 205 texts were added. The size of the sample after this second stage was therefore 705 texts.

The final stage of selecting texts was conducted after completing budgetary estimates of the number of texts that could be included in the final CJ rating task. We determined that a total of 1300 texts could be included in that task, and therefore added 595 further texts. These were added according to the three principles above, though increasing data sparsity meant that flexibility was required with regard to principles 2 and 3b.

L1	themes	count	mean word count
Urdu	6	16	585.0625
Punjabi	9	29	620.4828
Bulgarian	3	45	533.7556
Chinese	7	45	524.2667
Czech	8	45	644.1333
Dutch	13	45	590.1333
Finnish	11	45	621.6444
Greek	5	45	570.2444
Italian	13	45	566.8444
Korean	14	45	565.8444
Lithuanian	11	45	550.9111
Macedonian	11	45	624.5556
Norwegian	8	45	617.0667
Persian	3	45	614.6444
Polish	18	45	611.3778
Portuguese	9	45	542.2667
Russian	13	45	609.8
Serbian	10	45	584.7111
Swedish	11	45	567.9333
Tswana	5	45	501.7333
Turkish	7	45	646.7111
Hungarian	4	46	485.587
Japanese	5	46	507.7174
Spanish	15	100	630.05
French	13	102	639.3922
German	48	106	653.5472

Table 1: Text sample by L1 background of learners

A summary of the final sampling of texts is presented in Table 1. It is organised by L1 and shows that:

1. For 19 of the L1 groups, a total of 45 texts were selected. There were also 45 texts selected from the Pakistani sub-corpus, which is made up of Urdu and Punjabi texts. The higher number of texts from the Hungarian and Japanese sub-corpora resulted from making up the number of texts to the desired total of 1300. The higher numbers of Spanish, French, and German texts is due to the addition of the earlier sample of 222 manually annotated texts, which were mostly drawn from these sub-corpora;

2. There is some variance in the number of themes per L1 group, for the reasons outlined above;
3. There is some variance in mean word counts; some L1 groups wrote longer essays than others.

Comparative judgement assessment task

To generate proficiency scores for all texts, a comparative judgement task was conducted. The steps were as follows:

1. A group of trained raters were recruited by placing advertisements on language assessment-related electronic mailing lists. Applications were invited from participants who had the following qualifications:
    a. A minimum of 5 years' experience teaching ESL/EFL;
    b. A recognised qualification in Teaching English as a Foreign/Second Language or a related field (e.g. undergraduate or postgraduate degree in Applied Linguistics, Language Testing and Assessment, etc.; CELTA, DELTA);
    c. Training in L2 writing assessment;
    d. Experience of assessing L2 writing, preferably as an examiner;
    e. Private access to a personal computer with an internet connection.

    68 judges were selected from the resulting pool. All met the above criteria.
2. All texts were processed to ensure that:
    a. the correct prompt appeared as their title;
    b. their paragraphing was standardised, such that line breaks appeared between paragraphs (this was done to ensure that computational tools for assessing paragraph-based cohesion measures worked correctly);
    c. Any html tags and other artefacts from earlier text processing operations were removed;
3. All texts were uploaded to the No More Marking CJ platform (nomoremarking.com);
4. A task was constructed in which judges were asked to decide "Which of the two texts is a more advanced piece of second language writing?";
5. Judges were added to this task, with a quota of 200 comparisons each;
6. Judges completed their comparisons over a three-week period in April/May 2024. The task generated 13,163 total comparisons, at a rate of approximately 20 per text. Two judges dropped out during the task, leaving 66 remaining. Quotas were not adjusted to compensate for these lost comparisons, since early indications suggested that a high enough reliability (> .80) could be achieved without them;
7. The resulting data was processing in *R* using the Bradley-Terry model (using the *btm* function of the *sirt* package), which generates a rating scale assigning a probabilistic score to each item which denotes its quality relative to all other texts. Code for reproducing this model is included in the same package as this document;
    a. Data screening procedures identified two judges whose data did not meet acceptable standards. These judges misfit the overall CJ model, and additionally took a median of less than 40 seconds to complete each comparison. These judges and their data were removed from the study;
    b. The model was reconstructed without the two removed judges (i.e. based on a total of 12,765 comparisons made by 64 judges)

The scale separation reliability (SSR) of the final model was calculated to be SSR = .823. The correlation between the aforementioned rubric-based text ranks and the CJ ranks was rho = .768.

References

Thwaites, Peter; Charalambos, Kollias; Paraskevi, Kanistra; Paquot, Magali, 2024, “ICLE500", https://doi.org/10.14428/DVN/RIOSSC, Open Data @ UCLouvain, V1, UNF:6:F+Zi/uCo7OFbc+dOhfym8g== [fileUNF]