ICLE_500 metadata documentation

The ICLE_500.csv and ICLE_500.xlsx files provide basic text metadata and rubric-based assessment scores for 500 argumentative texts from the International Corpus of Learner English (ICLE; Granger et al., 2020). The texts themselves are also included in this data release. Extended metadata for the texts, including information on the conditions under which the texts were collected, are available via the ICLE website (https://corpora.uclouvain.be/cecl/icle/).

This document describes the sampling of the 500 texts from the corpus. The process of generating rubric-based assessments of the texts is outlined in the detailed report (Kanistra & Kollias, "Aligning the ICLE 500 written scripts with the CEFR: the technical report") included in this data release.

Text selection

The general goal in sampling these 500 texts was to be representative of the ICLE argumentative corpus as a whole. Three specific principles, each relating to a feature of the texts in the corpus, were chosen in order to guide this task:

1. Text length
    - The mean length of texts in the corpus was 598 words. To be representative of this, 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 be representative of the most frequently occurring 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 selected 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.

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

1. For 24 of the L1 groups, a total of 20 texts were selected. There were also 20 texts selected from the Pakistani sub-corpus, which is made up of Urdu and Punjabi texts.
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.

L1	n	prompts	mean_words
Urdu	4	2	587.25
Punjabi	16	4	625.9375
Bulgarian	20	3	540.05
Chinese	20	4	542.15
Czech	20	6	630.15
Dutch	20	5	568.15
Finnish	20	6	623.2
French	20	5	596.6
German	20	9	595.05
Greek	20	4	572.95
Hungarian	20	1	498.45
Italian	20	5	580.85
Japanese	20	2	526.45
Korean	20	7	556.4
Lithuanian	20	5	554.2
Macedonian	20	6	642.25
Norwegian	20	5	607.95
Persian	20	2	602.75
Polish	20	6	624.35
Portuguese	20	5	551.75
Russian	20	5	610.55
Serbian	20	6	581.25
Spanish	20	5	598.75
Swedish	20	4	570.7
Tswana	20	4	523.5
Turkish	20	3	634.65

Table 1: Text sample by L1 background of learners

Reference

Granger, S., M. Dupont, F. Meunier, H. Naets, & Paquot, M. (2020). *The International Corpus of Learner English* (version 3). Louvain-la-Neuve: Presses universitaires de Louvain.