This document contains a brief outline of the subcorpus and of the process used to create it. More details are available in the other documents.

The initial sample includes all texts from levels 1-15 in the original database, from the following nationalities: Brazilian, Chinese, Taiwanese, Russian, Saudi Arabian, Mexican, German, Italian, French, Japanese, and Turkish.

The texts in the dataset were then processed as follows:

- Common markup tags were cleaned from texts, and texts with large amount of varied markup tags were removed from the sample.
- Ultra-short texts (<20 words) were removed from the sample.
- Texts containing large amounts of non-English writing were removed from the sample.
- Texts containing large amounts of duplicate material that appeared in other texts were removed from the sample.
- Text with outlier wordcounts (i.e. extremely high or extremely low) were removed from the sample.
- The prompt that each text was written in response to was identified, and texts were divided into two separate datasets, based on whether they were written in response to the original prompt or the second one.

In addition, some additional variables were derived afterward from the original variables (e.g. CEFR level was derived from the proficiency level).

If you use the EFCAMDAT Cleaned Subcorpus in your work, please cite the following paper (a copy of which is available here as one of the documents):

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 220-236.

https://doi.org/10.1075/ijlcr.20009.sha