CORPUS REPORT

Refining and modifying the EFCAMDAT

Lessons from creating a new corpus from an existing large-scale English learner language database

Itamar Shatz
University of Cambridge

This report outlines the development of a new corpus, which was created by refining and modifying the largest open-access L2 English learner database – the EFCAMDAT. The extensive data-curation process, which can inform the development and use of other corpora, included procedures such as converting the database from XML to a tabular format, and removing problematic markup tags and non-English texts. The final dataset contains two corresponding samples, written by similar learners in response to different prompts, which represents a unique research opportunity when it comes to analyzing task effects and conducting replication studies. Overall, the resulting corpus contains ~406,000 texts in the first sample and ~317,000 texts in the second sample, written by learners representing diverse L1s and a large range of L2 proficiency levels.

Keywords: data curation, corpus cleaning, English as a second language, EFCAMDAT

1. Introduction

Learner corpora are increasingly being developed from data that originates in large-scale online platforms. This is beneficial, since the growing size of such corpora enables the analysis of large amounts of learner data, in ways that were not possible before (Callies, 2015; McEnery, Brezina, Gablasova, & Banerjee, 2019). However, with these new data sources come new challenges, which require new developments in terms of how researchers curate and analyze learner corpora. One notable challenge is the need to develop approaches to working with data that was originally collected with educational or social goals in mind, rather than research, since such data is often messy and requires substantial processing before

https://doi.org/10.1075/ijlcr.20009.sha
International Journal of Learner Corpus Research 6:2 (2020), pp. 220–236. issn 2215-1478 | e-issn 2215-1486

it can be properly analyzed. In addition, because of the increasing size of these new corpora, new approaches to data curation and analysis must be scalable, so they can be applied effectively on large-scale datasets, which puts an emphasis on the use of quantitative and NLP-based approaches.

The present report addresses this topic by discussing the development of a new derivative corpus from an existing learner language database. The goals of this report are both to introduce the new derivative source, and to explain how it was developed, in order to inform future data curation and analysis by researchers working with other learner corpora, and potentially with other corpora in general.

In particular, the original database used in this report is the *EF Cambridge Open Language Database* (EFCAMDAT), which is the largest open-access L2 English learner database, with 1,180,310 texts written by 174,743 learners from various nationalities (Geertzen, Alexopoulou, Baker, Hendriks, Jiang, & Korhonen, 2013; Huang, Geertzen, Baker, Korhonen, & Alexopoulou, 2017; Huang, Murakami, Alexopoulou, & Korhonen, 2018). The texts in the EFCAMDAT were submitted by learners to EF's online English school, which spans 16 English proficiency levels aligned with common proficiency standards such as the CEFR. Each level consists of 8 units, and upon completing a unit, learners are tasked with writing a text, which is then graded. If the learner receives a passing grade, they advance to the next unit; otherwise, they repeat the unit. The texts cover a variety of topics, such as reviewing a song for a website or describing one's favorite day.

The EFCAMDAT is pseudo-longitudinal overall, as learners generally complete only parts of the learning program. However, it contains substantial longitudinal data, since many learners complete sequences of tasks across increasing levels of proficiency, and researchers can track individual learners using the *learner ID* variable. In terms of metadata, the EFCAMDAT lists learners' English proficiency and their nationality, and learners were only added to the database if their nationality matched their country of residence (Alexopoulou, Michel, Murakami, & Meurers, 2017). Prior research on the EFCAMDAT used learners' nationality to estimate their L1, an approach that has been validated empirically (Alexopoulou et al., 2017; Huang et al., 2018; Murakami, 2013).

I developed the derivative version of the EFCAMDAT because I wanted to conduct a large-scale quantitative study of L2 lexical development, and found that I first needed to make several substantial modifications to the EFCAMDAT. As such, some of the decisions made in the course of creating the derivative corpus may not work well for other types of research. For example, the removal of duplicate texts described below may interfere with analyses that focus on formulaic language. However, researchers can choose to implement only some of the procedures that I outline in this report; to facilitate this, I make the relevant programmatic scripts available, together with partially cleaned versions of the new corpus.

Overall, the outcome of this data-curation process, in terms of the new corpus, led to significant modifications in three key areas:

1. **Format.** The new corpus is in a tabular format, rather than the EFCAMDAT's original XML format.
2. **Content.** The new corpus has been cleaned to remove texts containing issues that are likely to interfere with analyses relating to lexical development.
3. **Structure**. The new corpus is split into two samples, to account for some tasks containing groups of texts written in response to different prompts.

2. Preparing the new corpus

2.1 Selecting the sample

Because the new corpus was created with large-scale quantitative analyses of L2 lexical development in mind, the first step was to ensure that there were sufficient texts available for each combination of nationality and L2 proficiency level. Accordingly, I selected only those nationalities and proficiency levels that had enough texts for my analyses. This is in line with many prior studies that used the EFCAMDAT. For example, Murakami (2016) and Shatz (2019) examined only the top 10 nationalities with most texts in the EFCAMDAT, while Alexopoulou, Geertzen, Korhonen, and Meurers (2015) and Geertzen, Alexopoulou, and Korhonen (2014) examined only the top five.

In terms of proficiency level, texts from levels 1–15 were kept, while those from level 16 were omitted. There were relatively few texts at the omitted level (1,940, only 0.16% of the total), which were spread across multiple nationalities and tasks. In addition, levels 1–15 were grouped in bands of 3 based on EF's guidelines, while level 16 was on its own (Geertzen et al., 2013). Furthermore, level 16 was the only level listed as being above the maximum proficiency level set by several proficiency standards, such as the TOEFL.

In terms of nationality, texts from the 11 nationalities with most texts were kept: Brazilian, Chinese, Taiwanese, Russian, Saudi Arabian, Mexican, German, Italian, French, Japanese, and Turkish. These nationalities accounted for the vast majority of texts in the corpus (~89%), and there were relatively few texts spread out across the other 187 nationalities. Overall, 1,051,939 texts fit these criteria (89.1% of the texts in the EFCAMDAT).

2.2 Format: Converting from XML to a tabular format

The EFCAMDAT was originally made available in XML format; a sample text with the original XML formatting appears in Figure 1.

```
<writing id="135" level="2" unit="5">
        <learner id="73293" nationality="br"/>
        <topic id="13">Making notes for a visitor</topic>
        <date>2012-03-14 13:01:30.430</date>
        <grade>96</grade>
<text>
  Welcome to my house. Near the my house there is recreation center. Opposite to the
recreation center there is a soccer stadium. Between the recreation center and the soc-
cer stadium there is many restaurants. You guys enjoy!
</text>
</writing>
```

Figure 1. Sample text from the EFCAMDAT, with original XML formatting

The EFCAMDAT was imported from XML format using R, together with the *XML* package and a custom function (Lang, 2020). This converted the texts and all their metadata into tabular *xlsx* format, where each row represents a single text. In addition, the following markup tags were modified, to clean the texts for analysis:

- 762,221 <*br/*> and 35 <*br*> tags were replaced with a space.
- 88,343 *&amp;quot;* tags were replaced with a single set of quotation marks.
- 6,872 *&amp;* tags were replaced with the word *and*.
- 1 </*code*> tag was replaced with a space.

2.3 Content: Analyzing and removing texts

2.3.1 *Texts with problematic markup tags*

A small number of texts contained the *&lt;* and *&gt;* markup tags, which stand for ‘<’ and ‘>’ respectively. Texts containing these tags were removed, because they were generally accompanied by problematic data, such as improperly formatted error tags provided by teachers, together with suggested corrections. This included, for example:

- *&lt;&lt;&lt;&lt;IS&lt;correct&gt;/correct&gt;*
- *&lt;&lt;C, PU&lt;*
- *MY&lt;&lt;x&gt;y&lt;My*).

These tags were not supposed to be in the version of the EFCAMDAT used here, which is meant to be free of annotations, and they would have interfered with future analyses, for example by inserting words into the text that the learner did not write. The reason why the full texts were removed is that the tags were inconsistent in terms of structure, so there was no simple scalable way to remove them while preserving the original texts they were in.

There were 4,554 *&lt;* tags and 1,631 *&gt;* tags in the sample, spread across only 1,329 texts (0.1% of texts at this stage). To remove them, two R packages were used: *stringr* to detect the relevant strings in texts (Wickham & RStudio, 2019), and *dplyr* (Wickham, François, Henry, Müller, & RStudio, 2019) to filter texts based on the detected strings. After this removal, 1,050,610 texts remained.

2.3.2 *Ultra-short texts*

Ultra-short texts were defined as texts with fewer than 20 words, since such texts were below the minimal wordcount that learners were instructed to write, even at the lowest proficiency level. These texts often contained various issues. For example, many contained just random symbols (e.g. "???,??!??????,????????!" in text #876464) or only a few words (e.g. just "Hi," in text #613359). Similarly, there were over 20,000 such texts that were close variants of the same sentence ("Good evening. How are you. I'm fine, thanks. We're busy. Good night.").

Wordcounts were calculated using the *stringr* package (Wickham & RStudio, 2019) and a custom search pattern. 68,976 ultra-short texts were removed from the sample (6.6% of texts at this stage). Their average length was 13.49 words (*median* = 13, *standard deviation* = 3.38). Most of these texts (51,460, 74.6%) came from the first three tasks, with the majority (40,152) coming from the first one. After this removal, 981,634 texts remained.

2.3.3 *Non-English texts*

Texts that were not written in English were removed. This included texts that contained gibberish of various forms, texts that were written entirely in a foreign language, and texts that contained substantial portions written in a foreign language. This problematic material often appeared due to technical issues, such as when the L1 instructions were copied into the text.

These texts were identified using the *cld2* library in R, which relies on a Bayesian classifier that identifies the language of texts (Ooms, 2018). The threshold for removal was the maximal proportion of English in the text (0.99), to ensure that the texts did not contain substantial portions of foreign-language material. Overall, only 16,925 texts containing significant levels of non-English text were removed (1.7% of texts at this stage). After this removal, 964,709 texts remained.

2.3.4 *Duplicate texts*

Duplicate texts were texts that were almost identical to each other in substantial portions. This generally occurred as a result of reusing source material from the task almost verbatim. For example, the texts in task #64 often had the exact same opening in response to the prompt "Claiming back your security deposit": "Dear Sir, I am writing to ask your advice about a problem I have with my landlord and the real estate agent...".

As with the other steps in the cleanup process, there are advantages and disadvantages to this removal. Specifically, the main advantage of removing these texts is that the direct reuse of source material could obscure L1 effects and other linguistic patterns in unpredictable ways. Conversely, the main disadvantage of removing these texts is that this could lead to the removal of some meaningful linguistic patterns, such as the use of formulaic language. However, this concern was mitigated, as this issue appeared to be relatively task dependent, rather than proficiency dependent. For example, the task with the highest proportion of duplicate texts (69.7%) was task #64, which is relatively advanced. This suggests that the issue of duplicate texts occurred, to a substantial degree, as a result of task effects and idiosyncrasies in the learning situation. In addition, the potential issues with this removal were further mitigated, as texts were removed only when they contained a substantial portion of identical, overlapping phrasing, down to letters and punctuation marks.

To calculate similarity between texts in the database, the *stringdist* package in *R* was used (Van der Loo, 2014). Specifically, the analysis used the *hamming* method, an edit-based algorithm that calculates the number of substitutions required to get from one string to another. To use it, trimmed versions of each text were created, which contained only the first 100 characters, since this method requires that the compared texts be of identical length. Texts were trimmed specifically to 100 characters, as the shortest text was 104 characters, and 100 represented a close and round number. This is beneficial when determining the similarity threshold later, and provides a proportion that is simple to replicate.

Then, to determine the threshold of similarity at which texts would be considered duplicates, an initial analysis was conducted on a sample of texts from Brazilian and Japanese learners in tasks #1 and #73. This sample was chosen as it represents two distinctly different nationalities and tasks, which contain different numbers of texts (16,229 and 3,513 for Brazilian, 737 and 307 for Japanese, in tasks #1 and #73 respectively).

A similarity matrix was calculated for the texts in this sample, and duplicate texts based on a similarity threshold of '5' were extracted. This means that in cases where a trimmed text required fewer than five substitutions to be transformed into a different text in the sample, the two texts were designated as duplicates. Then, duplicate texts based on a similarity threshold of '10' were also extracted, and the results between the two thresholds were compared manually by examining the list of new texts that were identified as duplicates, and checking whether they appeared to include true duplicates or false positives. This process was repeated, each time increasing the threshold by increments of 5 (10→15→20…). Eventually, 40 was identified as the optimal threshold, since it appeared to lead to the identification of new duplicates compared to a lower threshold of 35, and because increasing the threshold to 45 appeared to lead to a substantial increase in false positives.

Finally, a similarity matrix was calculated on the main sample, using '40' as the threshold. Because each text must be compared against all other texts, this calculation involves potentially prohibitive computational complexity when run on large-scale datasets such as the EFCAMDAT. To resolve this, the analysis was run separately for each combination of nationality and task (for example, texts in task #1 among Japanese speakers, texts in task #1 among German speakers, etc.). This reduces the complexity of the calculation and is unlikely to have a substantial impact on its outcome, since within-nationality duplicates are more likely than between-nationality duplicates, and since between-task duplicates are unlikely.

Based on this, 194,722 texts were removed (20.2% of the sample at this stage). Certain tasks were more likely to contain duplicate texts; for example, 8.9% of texts in task #23 were removed as duplicates, compared to only 5.1% of texts in task #53. Higher proficiency tasks were less likely to have texts marked as duplicates, but there were many cases where higher-level tasks had a higher proportion of duplicates than lower-level tasks (the correlation between proportion of duplicates per task and task number was *Spearman's rho* = −0.46, $p < .001$). After this removal, 769,987 texts remained.

2.3.5 *Outlier texts based on wordcount*

This step targeted texts that were anomalously short or long. Such texts often suffered from various issues, such as the inclusion of large amounts of irrelevant material, for reasons that are unclear. For example, text #455618 was anomalously long, with 129 words at task #1 where the average wordcount was 32, and contained a letter about a company's logo in response to the prompt "Introducing yourself by email".

Outlier texts in terms of wordcount were identified using *Tukey's method*. This means that, for each task, outlier texts were those that had a wordcount 1.5 *interquartile ranges* (IQR) below the 1st quartile or above the 3rd quartile of wordcounts for texts from the same task (Kaliyaperumal, Kuppusamy, Arumugam, Kannan, Manoj, & Arumugam, 2015). Accordingly, a different set of problematic texts were identified using this method compared to the one for removing ultra-short texts, since this method accounts for differences in wordcounts between tasks. For example, this means that text #1211137 was removed in this step, since it had a wordcount of 24 at task #26, where the average wordcount was 63. Note that it would have been insufficient to use only this method without first removing ultra-short texts, because of the low average wordcount in many of the low-proficiency tasks, especially when ultra-short texts are included.

Based on this analysis, 34,607 texts were removed (4.50% of the sample at this stage). Of these, 5,717 (16.52%) were short outlier texts and 28,890 (83.48%) were long outlier texts. After this removal, 735,380 texts remained.

2.4 Structure: Classifying texts based on prompt

To explain this process, it helps to first define three terms:

- ***Task:*** this is the specific lesson that learners' texts are categorized under (e.g. "task #11"). Task numbers are listed sequentially in the EFCAMDAT, and range from 1–128, with 8 tasks per proficiency level.
- ***Prompt:*** this is the prompt that texts are written in response to (e.g. "Writing a weather guide for your city"). Each task has a corresponding prompt listed in the EFCAMDAT.
- ***Topic:*** this is the topic that a text revolves around (e.g. "weather"), as determined by classification software that will be described in this section.

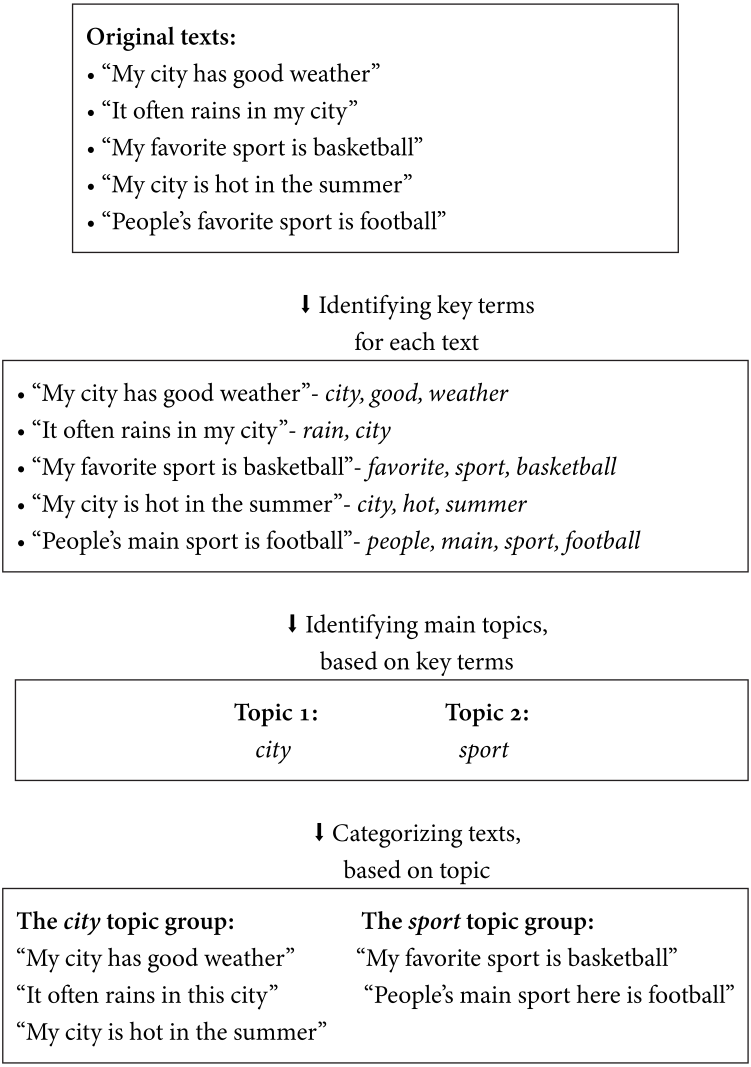
Many texts in the EFCAMDAT did not correspond to their listed task prompt. The reason for this issue was as follows:

- Initially, each task was associated with a certain prompt. For example, task #11 had the prompt titled "Writing a weather guide for your city".
- At some point, the prompts for some tasks were replaced with new ones. For example, the prompt for task #11 was changed to something such as "Describe people's favorite sport in your country".
- This change in prompt was *not* reflected in the database. Accordingly, all texts belonging to the same task number were listed together, regardless of which prompt they were written in response to. For example, task #11 contained texts written in response to the prompt on writing a weather guide, together with texts written in response to the new prompt on describing people's favorite sport.

Accordingly, it was necessary to do the following:

- Determine which tasks contained groups of texts corresponding to multiple prompts.
- Determine how many prompts were used in such tasks.
- Categorize the texts in such tasks based on the prompt that they corresponded to.

Since no information regarding the different *prompts* was available in the database, it was necessary to find a scalable way to analyze the *topics* that texts revolved around. To do this, I first grouped texts from each task (e.g. task #1, task #2…), and used the *tm* package in R (Feinerer & Hornik, 2018) to create a document-term matrix, with the term frequencies for each text. Then I used the *topicmodels* package (Grün & Hornik, 2011) to estimate a *latent Dirichlet allocation* (LDA) model using *Gibbs sampling*. For a visual representation of this process, see Figure 2.



Original texts:

- "My city has good weather"
- "It often rains in my city"
- "My favorite sport is basketball"
- "My city is hot in the summer"
- "People's favorite sport is football"

⬇ Identifying key terms
for each text

- "My city has good weather"- *city, good, weather*
- "It often rains in my city"- *rain, city*
- "My favorite sport is basketball"- *favorite, sport, basketball*
- "My city is hot in the summer"- *city, hot, summer*
- "People's main sport is football"- *people, main, sport, football*

⬇ Identifying main topics,
based on key terms

Topic 1: *city*

Topic 2: *sport*

⬇ Categorizing texts,
based on topic

The *city* topic group:
"My city has good weather"
"It often rains in this city"
"My city is hot in the summer"

The *sport* topic group:
"My favorite sport is basketball"
"People's main sport here is football"

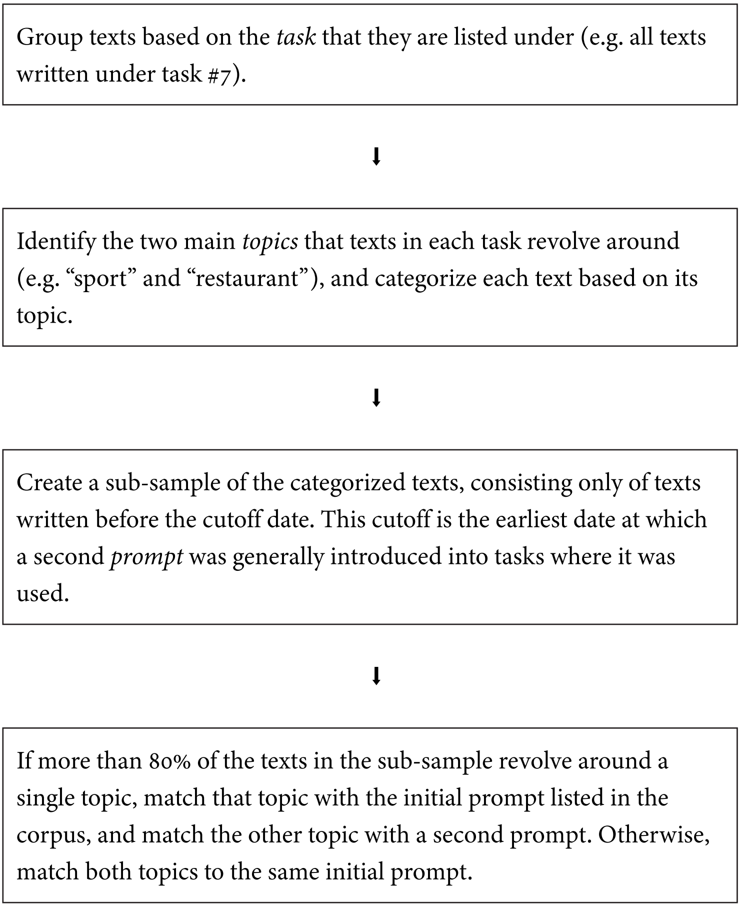
Figure 2. Rough illustration of the process used to classify texts based on topic

This process requires that the number of topics per task be specified in advance. Accordingly, to determine the appropriate number of topics, I started by running the process with two topics, and then tried increasing that number to three, while manually inspecting the texts. This revealed that the maximum number of prompts was '2', as dividing texts into more than two topics led to groupings that were *not* based on a difference in prompt. For example, if texts written in response to the prompts "a weather guide for your city" and "people's favorite sport in your country" were divided into more than two topics, then texts written in response to the same prompt would be separated; e.g. texts revolving around a weather guide might be split into those that primarily use keywords such as [*winter/cold/rain*] and those that use keywords such as [*summer/hot/sun*]. A single exception was task #13, where the classification software used the same keyword ('there') to classify texts from both topics. Accordingly, I re-ran the analysis for this task with three topics in the LDA model. I then examined the texts and combined two of the topics (under the keywords 'there' and 'house'), while the third topic (under 'neighborhood') was marked as corresponding to a different prompt.

Next, it was necessary to determine which tasks contained groups of texts corresponding to two prompts, and then classify texts accordingly. An examination showed that, in tasks with texts corresponding to two prompts, texts were initially written in response to the first prompt, until a certain date when the new prompt replaced the first. Accordingly, a sub-sample of the corpus was created, containing only texts submitted before 2012-07-04, which was established as the earliest approximate point when the second prompt was introduced. Then, the topics of the texts in the sub-sample were analyzed separately for each task:

- In cases where most texts (80%+) before the cutoff date belonged to a single topic, the task was categorized as having two prompts. Essentially, if most texts before the cutoff revolved around a single topic, this indicated that the topic corresponded to an initial prompt, while the less frequent topic corresponded to a second prompt that was introduced only after the cutoff. For example, if almost all of the texts before the cutoff revolved around the topic city, and almost none revolved around the topic sport, then it was likely that texts written about sport were based on a second prompt, which was introduced later.
- In cases where fewer than 80% of texts before the cutoff belonged to a single topic, the task was treated as having a single prompt. Essentially, if the texts before the cutoff date revolved around two topics in relatively similar proportions, then there was likely only one prompt for the task, since the similarity in proportion indicated that the division into topics was *not* based on a difference in prompt. For example, if texts before the cutoff revolved around the topics *restaurant* and *food* in relatively similar proportions, then it was likely that the texts were written in response to the same prompt, and that they simply used slightly different keywords.

One concern was that there might be tasks where one topic was much more common in the full sample. However, this was ruled out, given that the most extreme ratio between topics overall was 2.3:1 (between the second and the first topics in task #92), and the overall mean ratio between the first and second topics was 0.89 (*median* = 0.86, *SD* = 0.29). Conversely, the cutoff point used to determine whether two prompts were used was the much higher ratio of 4:1 (i.e. 80%). Overall, the procedure used to classify texts is outlined in Figure 3.



Group texts based on the *task* that they are listed under (e.g. all texts written under task #7).

Identify the two main *topics* that texts in each task revolve around (e.g. “sport” and “restaurant”), and categorize each text based on its topic.

Create a sub-sample of the categorized texts, consisting only of texts written before the cutoff date. This cutoff is the earliest date at which a second *prompt* was generally introduced into tasks where it was used.

If more than 80% of the texts in the sub-sample revolve around a single topic, match that topic with the initial prompt listed in the corpus, and match the other topic with a second prompt. Otherwise, match both topics to the same initial prompt.

Figure 3. Outline of the process that was used to identify and classify texts written in response to different prompts

In cases where all of the texts from a given task were established as having been written under the same prompt, they were all kept in the sample (31 tasks, 25.8% of total). Conversely, in cases where texts from a given task were established as having been written under two prompts, only texts written using the initial prompt were kept in the main sample (89 tasks, 74.2%). Accordingly, 329,318 texts (44.78% of texts) were designated as having been written in response to a second prompt, and were consequently separated into a second sample.

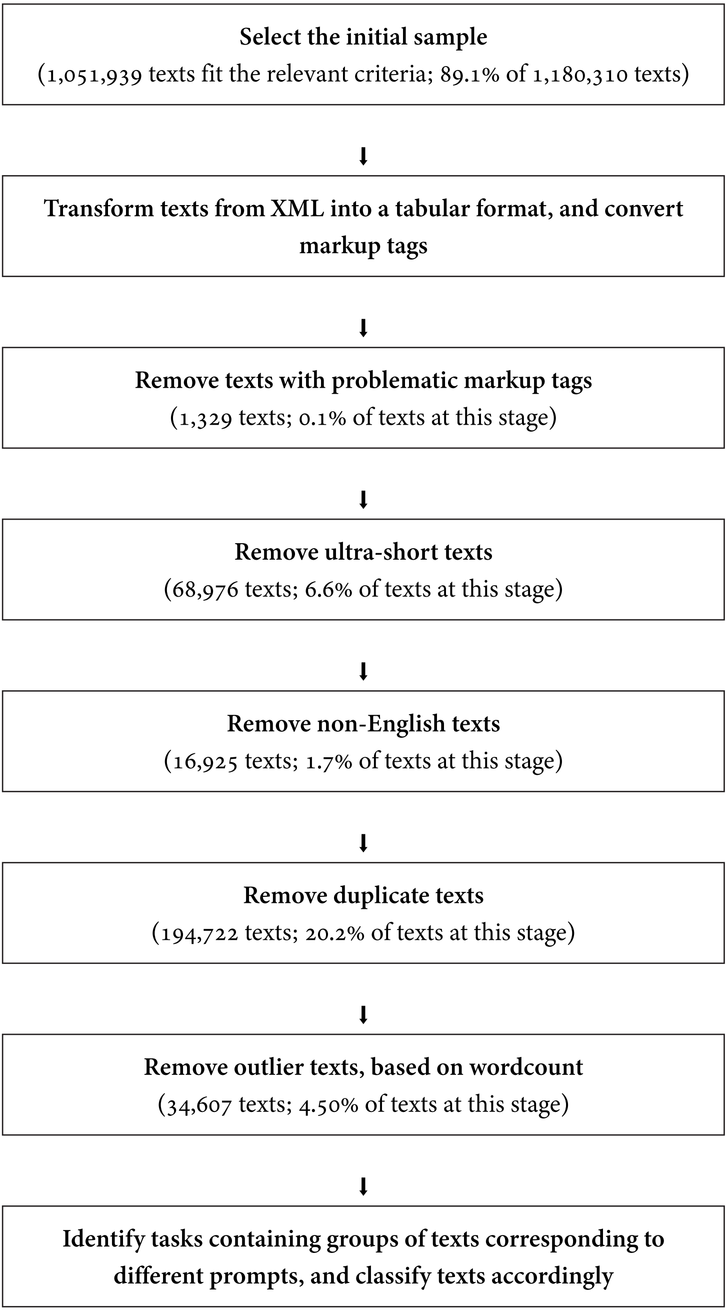
Finally, the texts in the second sample were further cleaned. This involved removing texts that were categorized as having been written in response to the second prompt despite being written before the cutoff date, which was the earliest point when the new prompt was generally introduced. The cutoff date used at this stage was 2013-04-03, which was later than the cutoff used previously. This is because the second prompt was often introduced around this later date, so using it allowed for the removal of more irrelevant texts. This led to the removal of 12,098 texts (3.67%), leaving 406,062 texts in the first sample and 317,220 texts in the second sample.

An important limitation of the second sample is that it does not list the prompts that learners responded to in their texts, since such data is not available in the EFCAMDAT. However, the original prompts from the first sample are still listed in the second sample, to maintain continuity between the samples; this ensures that the two samples share the same data structure, which means that researchers can easily concatenate them into a single sample if they wish. Nevertheless, it is possible to estimate the prompts manually, by reading the texts. Alternatively, it is possible to identify the key topics that the texts revolve around, by using the same keyword-extraction method that was implemented earlier; one such keyword is already listed for each text in the new version of the corpus, based on the earlier extraction process.

3. Discussion and conclusion

Overall, this report outlined a comprehensive process used to modify and refine a large-scale English learner database – the EFCAMDAT – in terms of its format, content, and structure. The process used to create the derivative corpus is outlined in Figure 4.

Based on this, from an initial database containing 1,180,310 texts, a corpus was created with 406,062 texts (~24,826,000 word tokens) in the first sample and 317,220 texts (~20,564,000 word tokens) in the second sample. These samples cover 120 and 89 topics respectively, and contain texts written by learners from 11 nationalities and with a large range of English proficiency levels (CEFR A1-C1). The numbers of texts per nationality and CEFR level in the new samples are listed in Table 1.



Select the initial sample
(1,051,939 texts fit the relevant criteria; 89.1% of 1,180,310 texts)

Transform texts from XML into a tabular format, and convert markup tags

Remove texts with problematic markup tags
(1,329 texts; 0.1% of texts at this stage)

Remove ultra-short texts
(68,976 texts; 6.6% of texts at this stage)

Remove non-English texts
(16,925 texts; 1.7% of texts at this stage)

Remove duplicate texts
(194,722 texts; 20.2% of texts at this stage)

Remove outlier texts, based on wordcount
(34,607 texts; 4.50% of texts at this stage)

Identify tasks containing groups of texts corresponding to different prompts, and classify texts accordingly

Figure 4. Summary of the preparation process of the corpus

Table 1. Number of texts in the derivative corpus, per nationality and CEFR level. Nationalities are listed by total number of texts in the first sample, in decreasing order

	Number of texts (first sample)						Number of texts (second sample)					
Nationality	**Total**	**A1**	**A2**	**B1**	**B2**	**C1**	**Total**	**A1**	**A2**	**B1**	**B2**	**C1**
Brazilian	149,297	75,497	45,407	20,989	5,830	1,574	164,241	85,191	42,105	25,520	9,412	2,013
Chinese	86,660	45,008	29,318	10,321	1,763	250	20,317	10,494	6,021	2,730	936	136
Mexican	34,559	19,296	9,847	4,102	1,114	200	30,204	15,998	7,645	4,500	1,740	321
Russian	32,243	12,295	10,885	6,329	2,066	668	17,078	7,249	4,652	3,449	1,443	285
German	24,705	8,041	7,860	5,051	2,698	1,055	16,717	4,652	4,487	4,083	2,669	826
French	19,135	7,626	6,253	3,688	1,242	326	13,384	4,610	3,755	3,188	1,528	303
Italian	18,959	5,899	6,832	4,291	1,466	471	16,469	5,046	5,010	4,166	1,749	498
Saudi Arabian	13,152	7,463	3,729	1,412	417	131	16,156	8,089	4,874	2,301	727	165
Taiwanese	11,711	4,116	4,298	2,506	650	141	10,900	3,668	3,731	2,490	893	118
Japanese	9,149	3,337	3,095	1,903	640	174	7,937	2,812	2,409	1,837	701	178
Turkish	6,492	3,085	2,067	914	301	125	3,817	1,683	1,064	769	253	48
Total	*406,062*	*191,663*	*129,591*	*61,506*	*18,187*	*5,115*	*317,220*	*149,492*	*85,753*	*55,033*	*22,051*	*4,891*

As noted earlier, the new corpus was created to facilitate large-scale quantitative analyses of L2 lexical development, using the data available in the EFCAMDAT. Accordingly, some of the procedures in the data-curation process may not be appropriate for other types of analyses; a notable example of this is the removal of duplicate texts, which could be an issue for analyses that focus on formulaic language. As such, to facilitate the use of the EFCAMDAT for other purposes, in addition to making the final version of the new corpus available, I have also made available additional versions of the corpus from different steps of the data-curation process, together with the key R scripts that I used. All these materials, together with other relevant ones, such as a glossary of variables, are available on the official EFCAMDAT site (https://corpus.mml.cam.ac.uk/). They are currently listed there under the "Resources" page, as the "EFCAMDAT Cleaned Subcorpus".

In addition to introducing the new derivative corpus, this report can also inform future work on other learner corpora, by identifying issues that researchers may encounter during data curation and analysis, and by proposing scalable solutions that they may use. This is something that is becoming increasingly necessary, given the growing use of large-scale learner corpora that are based on educational and social platforms, and that were therefore not originally collected with research in mind.

Acknowledgements

I wish to thank: my PhD supervisor, Dora Alexopoulou, for her continuous advice and support; Carlos Balhana and everyone else involved with the EFCAMDAT's technical aspects, for their work and assistance; EF Education First, for making their data openly available; Hughes Hall for their financial support; everyone who contributed to the software used in this research, for their important work; the editor and the anonymous reviewers, for taking the time to provide excellent comments.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1), 96–129. https://doi.org/10.1075/ijlcr.1.1.04ale

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180–208. https://doi.org/10.1111/lang.12232

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35–56). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

Feinerer, I., & Hornik, K. (2018). tm: Text Mining Package. Retrieved from https://cran.r-project.org/package=tm

Geertzen, J., Alexopoulou, T., Baker, R., Hendriks, H., Jiang, S., & Korhonen, A. (2013). *The EF Cambridge Open Language Database (EFCAMDAT). User Manual Part I: Written Production*. Retrieved from https://corpus.mml.cam.ac.uk/

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Millar, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, & A. Tseng (Eds.), *Selected proceedings of the 2012 Second Language Research Forum* (pp. 240–254). Somerville, MA: Cascadilla Proceedings Project.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Huang, Y., Geertzen, J., Baker, R., Korhonen, A., & Alexopoulou, T. (2017). *The EF Cambridge Open Language Database (EFCAMDAT): Information for users* (pp. 1–18). Retrieved from https://corpus.mml.cam.ac.uk/

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1), 28–54. https://doi.org/10.1075/ijcl.16080.hua

Kaliyaperumal, S. K., Kuppusamy, M., Arumugam, S., Kannan, K. S., Manoj, K., & Arumugam, S. (2015). Labeling methods for identifying outliers. *International Journal of Statistics and Systems*, 10(2), 231–238.

Lang, D. T.. (2020). XML: Tools for parsing and generating XML within R and S-Plus. Retrieved from https://cran.r-project.org/package=XML

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics*, 39, 74–92. https://doi.org/10.1017/S0267190519000096

Murakami, A. (2013). Individual variation and the role of L1 in the L2 development of English grammatical morphemes: Insights from learner corpora (Unpublished doctoral dissertation). Cambridge University.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4), 834–871. https://doi.org/10.1111/lang.12166

Ooms, J. (2018). cld2: Google's compact language detector 2 (Version 1.2). . Retrieved from https://cran.r-project.org/package=cld2

Shatz, I. (2019). How native language and L2 proficiency affect EFL learners' capitalisation abilities: A large-scale corpus study. *Corpora*, 14(2), 173–202. https://doi.org/10.3366/cor.2019.0168

Van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111–122. Retrieved from https://cran.r-project.org/package=stringdist

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2019). dplyr: A grammar of data manipulation. Retrieved from https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., & RStudio. (2019). stringr: Simple, consistent wrappers for common string operations. Retrieved from https://cran.r-project.org/web/packages/stringr/index.html

Address for correspondence

Itamar Shatz
University of Cambridge
Department of Theoretical and Applied Linguistics
Faculty of Modern and Medieval Languages
Raised Faculty Building, Sidgwick Avenue
Cambridge, CB3 9DA
United Kingdom
is442@cam.ac.uk