EF Cambridge Open Language Database (EFCAMDAT)

Geertzen, Jeroen; Alexopoulou, Theodora; Korhonen, Anna
Dept. of Theoretical and Applied Linguistics, University of Cambridge
jg532@cam.ac.uk; ta259@cam.ac.uk; alk23@cam.ac.uk

We present the EF Cambridge Open Language Database, henceforth, EFCAMDAT, a new open access database of written L2 English. EFCAMDAT was developed at the Dept of Theoretical and Applied Linguistics, at the University of Cambridge in collaboration with EF Education First, an international educational organisation. EFCAMDAT contains writings submitted to *Englishtown* the online school of EF, accessed daily by around 300,000 learners worldwide. The magnitude of EF operations has allowed us to build a resource of considerable size, currently containing 412K scripts from 76K learners summing up 32 million words. As new data come in, we expect to reach 100 million words by 2014 and be able to follow the longitudinal development of even more students.

EFCAMDAT consists of writings submitted to *Englishtown*, the online school of EF Education First, accessed by language learners all over the world (Education First, 2012). A full course in Englishtown spans 16 proficiency levels aligned with common standards such as TOEFL, IELTS and the Common European Framework of Reference for languages (CEFR). When students start a course at EF they are placed at the first level of a stage (levels 1, 4, 7, 10, 13, or 16) after a placement test and may proceed to higher levels through successful progression through coursework. Each of the 16 levels contains eight lessons, offering a variety of receptive and productive tasks. EFCAMDAT consists of scripts of writing tasks at the end of each lesson on topics like those listed in Table 1.

Table 1: Examples of essay topics at various levels. Level and unit number are separated by a colon.

ID	Essay topic	ID	Essay topic
1:1	Introducing yourself by email	7:1	Giving instructions to play a game
1:3	Writing an online profile	8:2	Reviewing a song for a website
2:1	Describing your favourite day	9:7	Writing an apology email
2:6	Telling someone what you're doing	11:1	Writing a movie review
2:8	Describing your family's eating habits	12:1	Turning down an invitation
3:1	Replying to a new penpal	13:4	Giving advice about budgeting
4:1	Writing about what you do	15:1	Covering a news story
6:4	Writing a resume	16:8	Researching a legendary creature

Given 16 proficiency levels and 8 units per level a learner who starts at the first level and completes all 16 proficiency levels would produce 128 different essays. Essays are graded by language teachers; learners may only proceed to the next level upon receiving a passing grade. Teachers provide feedback to learners using a basic set of error markup tags or through free comments on students' writing. Currently, EFCAMDAT contains teacher feedback for 36% of scripts.

The data collected for the first release of EFCAMDAT contain 551,036 scripts (with 2,897,788 sentences, and 32,980,407 word tokens) written by 84,864 learners. We currently have no information on the L1 backgrounds of learners, but metadata on the L1 background of learners is being collected for the second release of the database. Information on nationality is, thus, used as the closest approximation to L1 background. EFCAMDAT contains data from learners from 172 nationalities, with 28 nationalities having more than 100 learners, and 38 nationalities having more than 50 learners. Table 2 shows the spread of scripts across the nationalities with most learners.

Table 2: Percentage and number of scripts per nationality of learners

Nationality	Percentage of scripts	Number of Scripts
Brazilians	36.9%	187,286
Chinese	18.7%	96,843
Russians	8.5%	44,187
Mexicans	7.9%	41,115
Germans	5.6%	29,192
French	4.3%	22,146
Italians	4.0%	20,934
Saudi Arabians	3.3%	16,858
Taiwanese	2.6%	13,596
Japanese	2.1%	10,672

Most learners only complete portions of the program. Nevertheless, around a third of learners (around 28K) have completed 3 full levels, corresponding to a minimum of 24 scripts. Texts range from a list of words or a few short sentences to short narratives or articles. As learners become more proficient they tend to produce longer scripts. On average, scripts count 7 sentences (SD=3.8). Sample scripts are shown in the following figure.

1. LEARNER 18445817, LEVEL 1, UNIT 1, CHINESE

Hi! Anna,How are you? Thank you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english toghter! Bye! Anfeng.

2. LEARNER 19054879, LEVEL 2, UNIT 1, FRENCH

Hi, my name's Xavier. My favorite days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. LEARNER 19054879, LEVEL 8, UNIT 2, BRAZILIAN

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new prosperous city brought. I really enjoyed this song.

EFCAMDAT scripts have been annotated automatically with with Penn Treebank part-of-speech tags (Marcus et al., 1993) and grammatical relations according to the Stanford Dependency scheme (De Marneffe and Manning, 2008). Details of the automatic annotation and an evaluation of how these tools perform on learner data is presented in (Geertzen et al., 2012).

The database is accessed through a web based interface at http://corpus.mml.cam.ac.uk/efcamdat/. The interface supports selection of scripts from different proficiency levels and by learners of different nationalities, search for parts of speech and grammatical relations and export of raw text as well as tagged scripts. EFCAMDAT is freely available to the academic community subject to an end-user agreement protecting copyright.

References

De Marneffe, M. C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: Proc. of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Education First (2012). Englishtown. http://www.englishtown.com/.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2012). Automatic linguistic annotation of large scale l2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *in Proceedings of the 31st Second Language Research Forum (SLRF), Carnegie Mellon.* Cascadillla Press.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.