Working with Data from Real-World Corpora: A Case Study on Identifying Issues and Using Scalable Solutions

Itamar Shatz
University of Cambridge

The opportunity: new large-scale datasets

Language corpora are increasingly being based on data from social and educational online platforms, and the size of these new corpora allows researchers to analyze language use in ways that were not previously possible (Callies, 2015; McEnery et al., 2019).

The challenge: organize and clean large amounts of data

Online platforms generally do not collect data with linguistic research in mind, so their data is often “messy” or “dirty” in various ways.

Researchers must therefore develop new approaches for organizing and cleaning this data. Such approaches should generally be scalable, due to the size of these datasets, so they should rely primarily on quantitative and NLP-based techniques.

The solution: a case study

The original database. The *EF Cambridge Open Language Database* (EFCAMDAT) is the largest open-access L2 English learner database, containing ~1,180,000 texts written by ~174,000 learners from various nationalities (Geertzen et al., 2013; Huang et al., 2017, 2018). It contains data that was gathered from an online educational platform, which was run in many different countries over the course of years.

The process. Here, the original version of the EFCAMDAT was prepared for analyses of L2 lexical development, in terms of its *format*, *content*, and *structure*. This included the following:

The outcome. The resulting dataset contains two corpora, with ~406,000 texts (~24,826,000 word tokens) in the first and ~317,000 texts (~20,564,000 word tokens) in the second. The texts in the two corpora were written by similar learners in response to different prompts, which represents a unique research opportunity for studying task effects and conducting replications. The relevant data and code are available on the EFCAMDAT site (at <https://corpus.mml.cam.ac.uk/> under “Resources”).

Conclusion

New online platforms represent a potentially valuable source of language data, but researchers must develop scalable approaches to preparing this data before it can be analyzed. The case study outlined here provides insights into how this can be done.

Associated paper

This poster summarizes information found in: Shatz, I. (to appear). 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*.

References

- Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (Issue May, pp. 35–56). Cambridge University Press. <https://doi.org/10.1017/CBO9781139649414.003>
- Geertzen, J., Alexopoulou, T., Baker, R., Jiang, S., & Korhonen, A. (2013). *The EF Cambridge Open Language Database (EFCAMDAT) User Manual*.
- Huang, Y., Geertzen, J., Baker, R., Korhonen, A., & Alexopoulou, T. (2017). *The EF Cambridge Open Language Database (EFCAMDAT): Information for Users* (pp. 1–18). [https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000298](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000298).
- 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>
- 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>