

A call for more data-oriented reporting in Learner Corpus research

Nicolas, Lionel (Institute for Applied Linguistics, Eurac Research, Italy), Egon W. Stemle (Institute for Applied Linguistics, Eurac Research, Italy), Magali Paquot (Institut Langage et Communication, Université catholique de Louvain, Belgium), Hubert Naets (Institut Langage et Communication, Université catholique de Louvain, Belgium) and Alexander König (CLARIN ERIC, The Netherlands)

Scientific communication, including in Learner Corpus research (LCR), has traditionally prioritized the reporting of scientific findings and the methods used to derive them. In recent years, however, there have been several calls for placing greater emphasis on detailing how the data used to derive the findings was handled, especially in fields related to computer science (Mitchell et al., 2019; Gebru et al., 2018; Dodge et al., 2019). These evolving trends align with broader initiatives, driven in great part by institutional policies, aimed at enhancing research outputs by promoting best practices with respect to openness, reusability and reproducibility (among other aspects). The most prominent outcome from these initiatives is certainly the growing popularity of the FAIR principles (Wilkinson et al., 2016).

Our presentation will assess data-oriented reporting practices in LCR, regardless of the primary scientific objectives of the data, and will also detail facts and figures explaining, when reported, how data was collected and handled. In that perspective, we are currently surveying the most recent LCR literature published over the past 5 years, as referenced in the learner corpus bibliography³. Our survey relies on the coding of each publication for a specific set of features related to, at present, the reporting of formats, tools and licenses used.

Our preliminary findings suggest that progress can still be made. Building on these findings, we would like to open a discussion on future data-oriented reporting practices in LCR and contextualize our findings with respect to recent developments in Natural Language Processing, a computer-science-oriented research area with well-known ties to LCR. In particular, we will discuss the recent and noteworthy development of the “Responsible NLP Research Checklist”, which represents a promising and relevant basis to build upon for advancing data-reporting practices in LCR.

References

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
- Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004*.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.

³ <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html>