Type of Contribution: Full Paper

**Best Practices in Corpus Linguistics**

What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

Materials for this presentation are available at https://github.com/MartinSchweinberger/ICAME41\_BestPracticeInCorpusLinguistics

Schweinberger, Martin The University of Queensland, Australia *<*m.schweinberger@uq.edu.au>

*Abstract*

This talk represents a discussion note rather than a research presentation as it describes current and ongoing issues relating to Best Practices (BP) in Corpus Linguistics (CL). The presentation highlights problematic practices currently followed researchers engaged in CL, and proposes workable solutions which help guarantee transparency, replicability, and high quality of research outputs which will also serve to regain public trust. As such, the aim of this talk is to raise awareness about issues relating to BP in CL and to start a discussion of how to guarantee high quality of CL research.

While a discussion about BP, i.e. a method or procedure that is superior to alternatives because it produces results that are more reliable, transparent, replicable, or more ethical, in CL has recently begun (e.g. Berez-Kroeker et al. 2018; Ruhi et al. 2014), this discussion focuses almost exclusively on data compilation whereas issues relating to BP in data processing and analysis remain underexplored.

This paper argues that substantially more attention has to be placed on the repercussions of the Replication Crisis for CL and explores inferences that CL can draw from the Replication Crisis. The Replication Crisis is an ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s (Diener & Biswas-Diener 2019) which contributes to the public loss of trust that Humanities, Social Science, and Arts has experienced of the past two decades (McRae 2018, Yong 2018).

Current problems in CL that negatively affect the transparency of CL research are that analyses are often not reproducible and that the replication of existing research does not take place in a meaningful or substantive frequency.

It is argued that the reasons for these issues are that

- Replications is disincentivized

While journals would likely lose readership if similar studies were published repeatedly while researchers enact face-saving strategies as the publication of studies showing conflicting findings could negatively impact career trajectories

- Lack of resources

Even if researchers wanted to implement BP, support infrastructure including training and materials around Best Practices are not available to researchers to date

For individual researchers as well as teams, proposed solutions to these issues include:

- Following FAIR data requirements

Data should be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

- Publication of data

o clear examples for how to cite data should be provided by corpus compilers

o data should be published in online repositories

o Digital Object Identifiers (DOI) should be assigned to data before studies that are based on the data can be published

- Notebooks

R and Jupyter Notebook and their publication on publicly accessible repositories such as GitHib guarantee full transparency and replicability

- Scripts over Tools

The use of R or Python should be incentivized as scripts enable replication while the use of software applications (tools such as Sketch Engine, MonoConc, SPSS, etc.) which are often untransparent digital black boxes disincentivise replication to due restricted access and time restraints

- Documentation

Documentation of workflows, where to find what, and whom to ask for help are essential not only for on-boarding of new staff but also allow the evaluation of ongoing research practices

- Archiving

Completed projects (data, code, etc.) should be stored on publicly available repositories (e.g. GitHub) to allow public access, transparency, and replicability.

Given the highlighted issues, I would like to suggest that as a community, we should stress the importance of replicability and transparency of research by actively promoting replication in teaching contexts. In addition, journal editors as well as reviewers could ask for citation of data sources as well as requesting the submission of data and scripts when reviewing papers. Furthermore, more substantive investments in training and a support infrastructure for issues relating to data management and transparent data analysis options (R, Python, Git, Markdown, etc.) would be required to remedy existing problems. Finally, it is necessary to continue the discussion around BP in CL as well as to form research networks and facilitate panels that help highlighting issues related to BP in CL.

*References*

Berez-Kroeker, A. L., L. Gawne, S. S. Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. I. Beaver, S. Chelliah, S. Dubinsky, et al. 2018. “Reproducible research in linguistics: A position statement on data citation and attribution in our field.” *Linguistics* 56(1), 1–18.

Diener, E. and R. Biswas-Diener 2019. “The Replication Crisis in Psychology.” *NOBA Project* < https://nobaproject.com/modules/the-replication-crisis-in-psychology>.

McRae, M. 2018. “Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows.” *ScienceAlert* <https://www.sciencealert.com/replication-results-reproducibility-crisis-science-nature-journals>.

Ruhi, Ş., M. Haugh, and T. Schmidt 2014. *Best practices for spoken corpora in linguistic research*. Cambridge: Cambridge Scholars Publishing.

Yong, E. 2018. “Psychology’s Replication Crisis Is Running Out of Excuses. Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.” *The Atlantic* <https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/>.