Pre-conference workshop:

Exploring Powerful Tools to Ensure Robust and Reproducible Results in Corpus Linguistics

Convenors:

Martin Schweinberger (The University of Queensland, Australia),
Joseph Flanaghan (University of Helsinki),
Gerold Schneider (University of Zurich)

This workshop explores how powerful tools enable researchers to come up with efficient workflows and pipelines which allow them to stand on the shoulders of giants and at the same time produce robust and replicable results.

Fully scripted workflows, such as R or Python scripts have the advantage that results can be updated, reproduced, and shared at the push of a button (Flanagan 2017). Markdown and Text can be integrated to offer a seamless transition between data and publication. Powerful user-friendly tools such as LancsBox, LightSide, AntConc or VARD can be installed by every user to apply state-of-the-art approaches to obtain reproducible results in few, well-defined steps. The advantages of these tools are that they have the potential to counteract the increasing loss of public trust in research from the Humanities and Social Science (Yong 2018).

Advanced statistical approaches on the one hand offer new insights, easing the step from data to evidence (Suhr et al. 2019, Schneider et al. 2017), for example by offering higher levels of robustness, such as cross-validation, regulation, overfit warnings and built-in evaluation. But on the other hand, some may also introduce new challenges to reproducibility and robustness: Topic Modelling and sampling may depend on random seeds, seemingly similar parameters can lead to strongly different results across similar or even the same tool – how should we deal with them? As such, the workshop advances the discussion about Best Practices in (corpus) linguistics (Berez-Kroeker et al. 2018) and aims to raise awareness about existing resources and problematizes practices in (corpus) linguistics that hinder transparency, replicability, and high quality of research outputs.

The workshop proposes approaches and invites contributions that discuss issues relating to compiling, storing, handling, and of course analysing data according to best practices, which guarantee

transparency and high quality of research in (corpus) linguistics, as well as publication practices (pre-registration, open access, and pre-prints).

Specifically, the workshop addresses the following issues:

(i) How can the FAIR principles (Findable, Accessible, Interoperable, and Reusable) be observed?;

(ii) How can transparency and replicability of research be enhanced by using collaborative tools(e.g. Google Docs, Git, Docker, shiny R)?;

(iii) How can Jupyter or R Notebooks help to document analyses and making them available to the community and reviewers enable full reproducibility?;

(iv) How can researchers profit the most from integrating user-friendly out-of-the-shelf applications in their workflows?

(v) What are advantages and disadvantages of powerful blackbox methods (e.g. BERT) and or as opposed to simple close-to-text methods (e.g. Concordancing)?

(vi) Should we aim for robustness or reproducibility?

(vii) How can we document workflows and prevent data loss or corruption?

The workshop format will include detailed recommendations on practices and tools, presentations of submitted research contributions and a substantial open panel discussion. As such, the workshop will begin around 9 am and end at 4pm with breaks in-between.

Call for Papers

We particularly invite case studies on these above-mentioned issues. Abstracts should be between 400 and 500 words. Please send your abstract before January 31 to Martin Schweinberger (m.schweinberger@uq.edu.au). Notifications of acceptance will be sent out in mid-January.

Appendix

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, Edward and Biswas-Diener, Robert (2019). The Replication Crisis in Psychology. *NOBA Project* < https://nobaproject.com/modules/the-replication-crisis-in-psychology>.

Flanagan, Joseph. 2017. Reproducible research: Strategies, tools, and workflows. Studies in Variation, Contacts and Change in English, 19. < http://www.helsinki.fi/varieng/series/volumes/19/flanagan/ >.

Schneider, Gerold, El-Assady, Menna, and Lehmann, Hans Martin. 2017. Tools and Methods for Processing and Visualizing Large Corpora. Studies in Variation, Contacts and Change in English, 19. < http://www.helsinki.fi/varieng/series/volumes/19/schneider_el-assady_lehmann/>.

Suhr, Carla, Nevalainen, Terttu and Taavitsainen, Irma. 2019. *From data to evidence in English language research*. Leiden: Brill.

Yong, Ed (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/ >.