

Closing remarks and outlook

Stefan Th. Gries

University of California, Santa Barbara & Justus-Liebig-University Giessen

I am grateful for the opportunity the editors offered me to chime in a bit with regard to what's at the heart of this special issue: reproducibility, replicability, robustness, and generalizability, to use Flanagan's very useful terminology. Space does not permit me to outline things comprehensively so I will briefly allude to two to three questions/things from my – recurrent hedge coming up – probably much too narrow and subjective view of corpus linguistics.

First, **we need to diagnose whether there is a replication crisis in corpus linguistics, what its extent is, and what its causes are.** Is there one, and if there is one, is it due to honest analytical mistakes? to researchers not following best practices? to a lack of established best practices (and maybe also the transmission of these to new, younger generations of researchers)? Etc. At this point and to the best of my knowledge, I don't see one. While I have certainly read papers with whose analyses I vehemently disagreed, which sometimes led to papers criticizing methodological practices, in the research areas that I try to stay on top of, I cannot remember when I last read work that made me think 'oh, (more) evidence for the replication crisis'. I of course welcome us discussing reproducibility/replicability, etc. so we might avoid ever facing such a crisis and to develop fair and helpful best practices. However, I also prefer such discussions to be informed by what the actual state of affairs is rather than by what could be an uncritical adoption of the kind of crisis mode that may currently dominate other social sciences (whose findings are also paid much more attention by wider-read media, which may result in a greater need for sensational results?).

Second, when discussing solutions to whatever level of crisis there might be, **we need to be aware of the continuum of solutions to such a potential crisis.** Reproducibility/replicability can be dealt with differently stringent interventions: (i) sharing everything (any and all data and analytical code), (ii) sharing parts of the data/code, (iii) having methods sections that are detailed enough that studies can in principle be replicated, and of course more hybrid solutions. In reviewing and, less often, in reading published work, I find that even this last and least demanding standard is often not met. Obviously, the lowest-hanging fruit of (iii) should always be pursued but, maybe less obviously, solutions (i) and (ii) are not always possible or not always desirable, which brings us to a final, related issue.

Third, **our solutions need to fit our overall academic ecosystem,** or, more progressively, the ecosystem needs to be adapted. On the one hand, especially linguists working with specific kinds of communities know that full data sharing is often not possible, because, for instance, the community represented in a certain corpus might not agree to a full sharing of the data. On the other hand and maybe more widely applicably, full sharing also raises questions especially for early career researchers. Like it or not, vast parts of academia are captives of a publish-or-perish culture – do we really want to force the newly-minted Ph.D. graduate X to publish the complete data set of their dissertation at OSF while they're looking to incrementally publish more and more case studies out of it to build an academic career? This might lead to some well-funded lab Y with three postdocs and six doctoral students across some ocean downloading the data X collected painstakingly over the last three years and preempt much of the research X was still going to do. And, while I respect (and support!) the notion of giving more weight/academic credit to 'data work', let's be realistic: In the current academic ecosystem – at least in my current main habitat – X being cited for their corpus data by the next papers out of lab Y is going to be so much less valuable

for them getting grants and findings jobs than the publications they would have liked to do but now cannot do anymore because Y preempted them. Of course one might say, 'oh, but sharing also makes data available to junior scholars they might otherwise not have!' That is true and indeed a great consequence of greater openness in making data available, but do junior researchers have the same resources to utilize such data like a large, more established lab would have? Probably not. It's easy to be (too) idealistic) about data sharing etc. and I was once myself, till I tried to convert someone not yet tenured into sharing the corpus data they had collected but refused to make available for precisely these reasons, which made me appreciate that their individual concerns were very justified and it felt self-righteous to continue to insist. In current academia, with its current belief and value systems and with its current oversupply of graduates/researchers and underdemand (fewer jobs for them), full data sharing is more complicated and individually fraught and risky than lofty declarations may make one believe.

Again, I could only scratch the surface, but I do think we need a more precise diagnosis of the current state of affairs and especially a more comprehensive view of the realities on the ground before we jump into actions that potentially penalize the most junior ones.