INTRODUCTION

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 is at the heart of this special issue: reproducibility, replicability, robustness, and generalizability, to use Flanagan's 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 practices 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 has sometimes led to papers criticizing methodological practices, in the research areas that I try to stay on top of, I cannot remember when I have last read work that made me think: "Oh, (more) evidence for the replication crisis". I, of course, welcome discussions around reproducibility, replicability, etc., so as to 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 and replicability can be dealt with through differently stringent interventions: (i) sharing everything (any and all data and analytical code), (ii) sharing parts of the data or 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

https://doi.org/10.1075/ijcl.24120.gri | Published online: 10 June 2025
International Journal of Corpus Linguistics ISSN 1384-6655 | E-ISSN 1569-9811

lowest-hanging fruit – (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 within our overall academic ecosystem**; or, more progressively, the ecosystem itself 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 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 dataset from their dissertation on 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 that X painstakingly collected over the last three years and preempting much of the research X was still going to do. And, while I respect (and support!) the notion of giving more weight and 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 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 who might otherwise not have it!" That is true and indeed a positive consequence of greater openness in making data available, but do junior researchers have the same resources to utilize such data as a large, more established lab would? Probably not. It's easy to be (too) idealistic about data sharing etc., and I was once myself, till I tried to convince someone not yet tenured to share the corpus data they had collected but had 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 oversupply of graduates and 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.

Address for correspondence

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
United States

stgries@linguistics.ucsb.edu
https://orcid.org/0000-0002-6497-3958

Publication history

Date received: 26 July 2024
Date accepted: 19 August 2024
Published online: 10 June 2025