Commentary

Lauren Gawne*, Helene N. Andreassen, Lindsay Ferrara and Andrea L. Berez-Kroeker

Open research requires open mindedness: commentary on “Replication and methodological robustness in quantitative typology” by Becker and Guzmán Naranjo

https://doi.org/10.1515/lingty-2025-0018
Received February 18, 2025; accepted May 13, 2025; published online July 24, 2025

Becker and Guzmán Naranjo’s (2025) (henceforth B&GN) exploration of new methods for analysing existing typological data is a great example of the importance of transparency in research. This includes both transparency in regard to the data analysed as well as transparency in the methods of analysis. This open approach to data and methods requires open mindedness, on behalf of the current researchers, those whose work is being built on, and the reviewers and editors who facilitate this research.[1]

We come to this article from our experience as co-chairs of the Linguistic Data Interest Group (LDIG) of the Research Data Alliance, an international community working towards the technical and social changes needed to facilitate open research. Helene, Lauren and Andrea were founding co-chairs, with Helene and Lindsay

[1] We acknowledge that linguists work with many communities for whom Western paradigms of research, including Open Access data, are not appropriate ways to represent their relationship to language and linguistic data (cf. Holton et al. 2022). In general, we advocate for a shift towards “transparency”, where data and methods are open, and if they are not open the researcher is transparent in explaining why this is the case. Since B&GN are working with existing Open Access data, we do not discuss this distinction in any more detail, but we feel it is important to note that data can contribute to an advancement in open research in different ways.

***Corresponding author: Lauren Gawne [lɔ.ɹən gɒn]**, La Trobe University, Melbourne, Australia, E-mail: l.gawne@latrobe.edu.au
Helene N. Andreassen [hɛ'le:nɛ ɛn ɑn'dre:ɑsn̩], UiT The Arctic University of Norway, Tromsø, Norway
Lindsay Ferrara [lɪnzi fɛɹaɹə], Norwegian University of Science and Technology, Trondheim, Norway. https://orcid.org/0000-0003-3679-3404
Andrea L. Berez-Kroeker [ændɹea beɹez kɹoʊkə], University of Hawai‘i at Mānoa, Honolulu, USA

currently serving in this role. We work with linguists from a wide range of subfields to understand the challenges in implementing more transparent ways of working, and to improve open practices for data and research methods. Linguistics, as a discipline, still has quite a way to go in working with the kind of transparency that typifies both B&GN's work and the scholarship they build on. Berez-Kroeker and colleagues (2017) surveyed 270 articles from nine top international linguistics journals from a ten-year span between 2003 and 2012. Very few authors in these journals met survey metrics for basic transparency of data and methodology. Similarly, Gawne and colleagues (2017a) surveyed 100 descriptive grammars from the same period and found that very few authors made their data or methods explicit in the grammar. Data transparency in this genre should be a cause for concern to the audience of this journal, given that descriptive grammars are a primary source of data for linguistic typologists. Gawne et al. (2017b) replicated the methodology of Berez-Kroeker et al. (2017) with 50 papers from a five year period of *Linguistic Typology* (2012–2017, vol. 16.3–21.2). Only five of those papers drew on a publicly available corpus, meaning there are few studies that can be subject to the kind of scrutiny in B&GN's work. Three of the 50 papers in the survey used data that was unpublished or unknown in source, posing very high barriers to any potential for replication.

When we talk to linguists across subfields, we find that usually the challenges are as much about mindset as any specific technical barriers. Of course, current institutional incentives do not often encourage researchers to publish data, but for some, existing workflows make changing practice difficult. For others there is a feeling of vulnerability and potential new fronts of criticism that are roadblocks to changing practice. Within this context, however, there are researchers thinking critically about how to collect, manage and share their data as part of their research; Berez-Kroeker et al. (2022) is a handbook with 56 chapters that show how linguistics from a range of subfields, including linguistic typology, are managing linguistic data, with a focus on transparency. Researchers like B&GN, who are revisiting existing work with new analyses, are helping to promote the new normal of how we use and talk about open data.

B&GN use the term "replication" to discuss mechanisms that allow third-party verification of research results, discussing different elements of linguistic typological work that can be replicated. By invoking replication, the authors are drawing a direct link to the long-established scientific principle that others should be able to replicate your results and come to the same conclusion. In many empirical disciplines, it is possible to replicate an existing finding using new data. This is true for some subfields of linguistics; it should be possible to replicate an acoustic analysis of vowels in Australian English with a new group of Australian English speakers. For other subfields of linguistics, replicating an analysis with a new sample may not yield the same results; in Discourse Analysis we do not expect different people – or even the same person! – to tell a story in exactly the same way each time. In these contexts, rather than REPLICATE the original methodology with new data, it is possible to still retain a commitment to open data by sharing the original data (e.g. narrative recording or transcript) allowing others to REPRODUCE the analysis of the original piece of data.[2] This is a small, but important, distinction, because it centralises the importance of including data alongside analysis. In a subfield of linguistics such as linguistic typology, building a sample is a carefully considered part of the methodology, and collecting a different sample can lead to different findings. The research under discussion is a great illustration of the value of reproduction. B&GN focus on variation in outcomes by reproducing analysis on existing data with new methods. For research to be replicable, transparency in methods is required so others can follow those methods with new data. For research to be reproducible, transparency in both the method and the data is required, which widens the scope of open research. Therefore we find it useful to distinguish between REPRODUCTION of analysis using original data and REPLICATION of methods on new data.

In our work helping researchers think about reproduction and replication in their own research, the LDIG has created two main outputs that provide a scaffold to help researchers think about open data. The first is the Austin Principles of Data Citation in Linguistics[3] (Berez-Kroeker et al. 2018b), an annotated version of the FORCE11 Joint Declaration of Data Citation Principles, translating these principles into the context of linguistics. The Austin Principles centre linguistic data in the production and dissemination of linguistic research. The second output of the LDIG is the Tromsø Recommendations for Citation of Research Data in Linguistics[4] (Andreassen et al. 2019), which provides guidance for citing linguistic data in publications to researchers, academic publishers and resource providers. Together, these documents provide the reasoning for, and principles for the implementation of, more transparent practices regarding the presentation of linguistic data in research.

Of course, individual researchers and their practices are shaped by the larger research environment. Part of the work of LDIG has been to work with journals and publishers to encourage them to take up these data citation standards for published research. This helps normalise the practice. We hope that the discussion in the pages of *Linguistic Typology* provoked by B&GN's work will encourage the editorial board to adopt the Tromsø Recommendations and improve data citation norms for this journal in line with others such as *Glossa*, *Language Documentation & Conservation* and *Intercultural Pragmatics*. This would expand the current De Gruyter Mouton's data sharing policy used by the journal, where authors are encouraged to provide a

[2] See Berez-Kroeker et al. (2018a) for a more detailed discussion of replication in linguistics and the social sciences.

[3] https://site.uit.no/linguisticsdatacitation/austinprinciples/ Accessed 31 Jan 2025.

[4] https://zenodo.org/records/3672840 Accessed 31 Jan 2025.

data availability statement. Citation to underlying data encourages greater transparency within the scope of published work. The current phase of work for the LDIG involves training researchers, publishers, and research communities in linguistics in new ways of working that may be challenging for the individual but advance the development of the whole discipline. If you're interested in helping advance the conversation about linguistic data, LDIG is an open RDA Interest Group that welcomes new members.

B&GN's work shows that research reproduction can advance the broader discipline as methods are tested and refined. Many of the results from the original studies were confirmed in the work outlined in the target paper, reminding us that reproduction does not need to be a combative exercise. Results that were challenged open up potential new lines of inquiry. More important than any specific method on show, linguistic analysis is a larger enterprise than any single scholar or team of researchers, and we appreciate that B&GN have used this exercise to highlight a number of ways that methodological robustness can be better executed in linguistic typology. In particular, we appreciate the way they highlight that this is not just about making a dataset publicly available, but about openness and transparency across the whole workflow, from primary data to sampling, annotation and statistical analysis. Gawne and Berez-Kroeker (2018) outline some of the institutional barriers to recognition for this work; we are heartened to see that *Linguistic Typology* has not only published this paper, but made it the topic of commentary, continuing to normalise reproduction research as a valued academic activity.

We thank B&GN for their work, but we also thank Dryer (2018), Seržant (2021), Shcherbakova et al. (2023) and Berg (2020), for doing work that could be reproduced (a benchmark much scholarship falls short on) and subsequently having their work scrutinised and evaluated in this way. As B&GN note, "[t]here is no specific reason for choosing these papers other than the fact that the authors made their datasets available". To work in an open and transparent manner is to open yourself to critical evaluation. Linguistic typology advances because of researchers who have created accessible data as part of their work. This includes individual researchers working on specific languages, whose data is the basis of typological work, as well as those typologists who share the databases of their work. Open ways of working require open mindedness from the whole research community.

References

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Phillip Conzett, Christopher Cox, Koenraad D. Smedt, Bradley McDonnell & the Research Data Alliance Linguistic Data Interest Group.

2019. *Tromsø recommendations for citation of research data in linguistics* (Version 1). Research Data Alliance. Available at: https://doi.org/10.15497/RDA00040.

Becker, Laura & Matías Guzmán Naranjo. 2025. Replication and methodological robustness in quantitative typology. *Linguistic Typology* 29(3). 463–505.

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003–2012. Available at: https://sites.google.com/a/hawaii.edu/data-citation/survey.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018a. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group & The Linguistics Data Interest Group. 2018b. *The Austin principles of data citation in linguistics*. Version 1.0. https://site.uit.no/linguisticsdatacitation/austinprinciples/ (accessed 31 January 2025).

Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.). 2022. *The open handbook of linguistic data management*. Cambridge, MA: MIT Press.

Berg, Thomas. 2020. Nominal and pronominal gender: Putting Greenberg's Universal 43 to the test. *Language Typology and Universals* 73(4). 525–574.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798–833.

Gawne, Lauren & Andrea Berez-Kroeker. 2018. Reflections on reproducible research. In Bradley McDonnell, Andrea Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998*, 22–32. Honolulu: University of Hawai'i Press.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017a. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189.

Gawne, Lauren, Andrea L. Berez-Kroeker & Helene N. Andreassen. 2017b. Data citation in linguistic typology: Working towards a data citation standard in linguistics (poster). *Association for Linguistic Typology* 12. Canberra: December 11-15.

Holton, Gary, Wesley Y. Leonard & Peter L. Pulsifer. 2022. Indigenous peoples, ethics, and linguistic data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open Handbook of linguistic data management*, 49–60. Cambridge, MA: The MIT Press.

Seržant, Ilja. 2021. Slavic morphosyntax is primarily determined by its geographic location and contact configuration. *Scando-Slavica* 67(1). 65–90.

Shcherbakova, Olena, Volker Gast, Damián Blasi, Hedvig Skirgård, Russell Gray & Simon Greenhill. 2023. A quantitative global test of the complexity trade-off hypothesis: The case of nominal and verbal grammatical marking. *Linguistics Vanguard* 9(s1). 155–167.