Matti Miestamo* and Kaius Sinnemäki

Sampling matters: commentary on “Replication and methodological robustness in quantitative typology” by Becker and Guzmán Naranjo

https://doi.org/10.1515/lingty-2025-0040
Received April 11, 2025; accepted May 15, 2025; published online July 25, 2025

1 Introduction

The target article by Laura Becker and Matías Guzmán Naranjo (2025) (henceforth B&GN) discusses replication in language typology with a focus on the role of statistical modelling as part of the replication process. Overall, their article draws much needed attention to replication in the field and offers some guidelines for replication in typological research. In this commentary we focus on a couple of issues that the article raises directly or indirectly which we think merit further discussion or that we do not completely agree with. In this introduction we briefly comment on the notion of robustness. We then discuss transparency in Section 2, sampling in Section 3, and the statistical model in Section 4 before brief concluding remarks in Section 5. The main point we wish to make in this commentary relates to sampling and in that sense Section 3 forms the core of our contribution, but issues related to sampling are addressed in the other sections as well.

Regarding robustness, B&GN state the following: “Applied to typology, robust findings then need to hold across (i) different language samples, (ii) alternative linguistic analyses and annotations as well across (iii) different statistical methods.” We find it easy to agree with points i and iii, but it remains somewhat unclear how point ii is to be understood. They give Nichols et al. (2006) as the only example they know of a study that tests replicability across different ways of categorizing the data, pointing out that the authors of that paper “show that their findings on the distribution of morphological complexity remain similar when using inflectional, derivational as well as lexical inflectional metrics.” To us this sounds more like looking at

***Corresponding author: Matti Miestamo ['mat:i 'miestamo]**, University of Helsinki, Helsinki, Finland, E-mail: matti.miestamo@helsinki.fi. https://orcid.org/0000-0001-9801-9030
Kaius Sinnemäki ['kaius 'sin:emæki], University of Helsinki, Helsinki, Finland. https://orcid.org/0000-0002-6972-5216

different, even if closely related, linguistic phenomena, different aspects of (morphological) complexity, rather than just applying different analyses/annotations to the same data. Looking at the distribution of morphological complexity using inflectional, derivational or lexical inflectional metrics means testing a theoretical claim through the lens of different linguistic phenomena using different types of data, which is of course a very good scientific practice, but we are not sure whether it can be subsumed under replication. If the same data are analysed in different ways, using different principles of classification, we will necessarily get different results in terms of cross-linguistic frequencies and areal distributions – or rather incommensurable results as the types in the classification are different to begin with. We may be able to address a theoretical question or test a hypothesis using different categorizations of even the same data, but how that counts as replicating results in typology, remains unclear.

2 Transparency and replicability

The “Guidelines for better replicability in typology” that the authors give in their appendix are very welcome. What they say in those guidelines regarding explicating the language sample, the linguistic analysis and annotation as well as the code for statistical analysis, should be covered by any methodology guidebook or typology textbook in the chapter(s) explaining how typological studies are conducted, but unfortunately this rarely happens. Following these principles should be the default in our field, but it seems that, as the authors also point out, most typological studies follow them only to a certain degree or not at all.

Especially in older typological studies, it is not uncommon that the principles of building the sample are not mentioned or that not even the sample languages are listed. This may happen even in relatively recent studies, one example being Aikhenvald (2010) in which the author states that a high number of languages, approximately 700, from various languages families and areas have been examined, with the aim of including all languages that the author was able to find reliable information for. However, no listing of the languages is provided (other than the language index that lists those languages that have been mentioned in the book). The author argues that typologists should look at all languages for which they can find sources rather than working with samples (p. 12, see also Aikhenvald and Dixon 2017). Of course, including every language one is able to find data for, thus covering the sampling frame in full, is very good for the aim of discovering the full range of cross-linguistic variety, but this is highly resource-intensive and not feasible in most cases. This approach amounts to a maximal variety sample, and thus no stratification is needed, but then a minimal requirement should be to at least list all the languages in that sample with their genealogical and areal affiliations and, crucially, the sources used for each language. Obviously, in a study where the languages examined have not even been listed, no systematic information is given on the analysis and annotation decisions for individual languages or structures, i.e. which type in the proposed typology each case is assigned to; illustrative examples from a number of languages may be discussed and the generalizations may be stated, but there is no table or appendix giving the reader the details for each language.

Only if the languages are listed and the analysis and annotation decisions are given for all of them (as most studies fortunately do), can we ask the question how transparently these decisions have been argued for, i.e. how well the recommendations in the target article have been followed. In addition to listing the annotation decisions for each language, does the study only give a selection of illustrative examples and state the generalizations, or is evidence given for the analysis and annotation decision for each sample language?

Typological studies fare much worse in this regard than in the explicitness of sampling. As B&GN point out, relatively few studies give evidence for the analysis of each language examined. However, the picture that they draw is perhaps even too negative. They list four studies that "include at least one example for each of the annotation decisions made in the paper", saying that this is an exhaustive list as far as they know. We can add some studies to the list. The works by Stassen (1997) on intransitive predication and Stassen (2009) on predicative possession discuss examples of all sample languages, although the discussion is integrated into the chapters of the books instead of being given in alphabetized appendices. Miestamo's (2003, 2005) work on negation includes an appendix that contains examples for each sample language. Similarly, Keinänen's (2025) work on the functional extensions of evidentials explicates the analytic decisions for all sample language in an appendix. Moreover, when the *World Atlas of Language Structures* (WALS; Haspelmath et al. 2005) was being prepared around the turn of the millennium, the editors were consciously aiming at transparency as all authors were asked to provide examples for all languages included in their respective chapters. At least some 20 chapters follow this instruction and give examples for most sample languages to back up their annotation decisions (although typically not providing examples for languages whose feature value is of the form "feature X absent" or "no feature X"). For obvious reasons, the examples are not there in the printed atlas, but they are available in the electronic version distributed on the accompanying CD ROM and in the current online version.[1]

[1] It seems, however, that there is data missing in the current online version and only some of the examples present in the original 2005 CD ROM version can be found in the online version (Dryer and Haspelmath 2013).

3 On the role of sampling in (quantitative) language typology

A central argument of B&GN seems to be that the days of controlling bias via sampling are over because statistical modelling provides a more accurate alternative to controlling phylogenetic and contact-related biases.

Sampling has played an important role in typological research as the main tool to ensure that the results are generalizable. Biases that may be due to genealogical affiliation or language contact have mostly been addressed via (stratified) probability sampling (e.g., Tomlin 1986; Perkins 1989) or some form of stratified random sampling at the level of languages, genera, or families (e.g., Sinnemäki 2011; see also Widmann and Bakker 2006 who compare results obtained via a number of alternative sampling methods, including non-stratified random sampling). The idea in these approaches is that the sampling technique addresses the biases that depend on genealogical affiliation and/or language contact, and then inferential statistical methods, such as correlation tests, can be applied to assess the relation between the variables of interest. Variety sampling is another important type of sampling method in typology, but its function is to ensure representativeness when exploring linguistic diversity rather than to serve as a tool to address biases for statistical testing (e.g., Rijkhoff et al. 1993; Rijkhoff and Bakker 1998; Miestamo et al. 2016). A crucial difference between probability sampling and variety sampling relevant to the present discussion concerns sample size: probability samples constructed to address biases tend to have maximally 100 languages (cf. Perkins 1989), but variety sampling techniques allow for much larger samples as they do not need to worry about the independence of the sample languages to the same degree (see, e.g., Rijkhoff and Bakker 1998; Miestamo et al. 2016).

B&GN convincingly argue that statistical bias control is more efficiently and adequately addressed by statistical modelling than by sampling. In statistical modelling the confounding factors, such as genealogical affiliation and language contact, can be built into the model, and their effect can then also be assessed separately, as suggested already in earlier work (e.g., Bickel 2015). The proposed approach builds a Bayesian regression model that incorporates phylogenetic trees to address biases arising from genealogical affiliation and a Gaussian process for distances between language populations to address biases arising from language contact. It is easy to agree with B&GN that doing so makes it possible to assess the dependencies between languages in a more accurate way compared to addressing biases merely via sampling.

But if statistical modelling is so much more efficient and accurate in addressing statistical bias, is there any need to continue being mindful about sampling in typology? B&GN seem to claim that there is not. They contrast two approaches: (i) using established sampling methods to control for biases and then assessing the relationship between the variables of interest via statistical tests versus (ii) assessing the relationships between the variables of interest via statistical modelling and then no attention needs to be paid to sampling, that is, convenience sampling will do. However, we wish to argue for the importance of sampling even when using statistical modelling.

Our impression is that while approach (ii) resembles recent practices in the field, it is more radical than current practices. Since the 2000s two important developments have changed quantitative language typology. First, a dynamic approach to typology (cf. Greenberg 1978) was implemented that approaches language universals as diachronic pressures to type-shift (e.g., Maslova 2000; Bickel 2013). Second, multiple regression methods, followed by mixed effects modelling, were introduced to typology (e.g., Cysouw 2010; Jaeger et al. 2011).[2] As a result, typologists started to argue that instead of controlling bias via sampling, a better way would be to control it via building genealogical affiliation and contact effects into regression models (e.g., Bickel 2015). And since regression models require more data and allow one to build dependencies between sampling units into the model, typologists also began to move away from probability sampling in which the independence of sampling units is prioritized over sample size. These developments already shifted attention away from sampling which resulted in a somewhat uncomfortable lack of discussion on the role of sampling in quantitative language typology since early 2010s.

B&GN explicitly argue that sampling is no longer needed for bias control, but that leaves them in an awkward position, defending convenience sampling. We would like to suggest that there is a third and a more appropriate way for quantitative typology. Our proposal is built on the fact that sampling has more than one function, namely, (i) representing variation in the population and (ii) bias control. The idea of representativity is that to produce reasonable generalizations about the population, the sample needs to represent the variation in the population to a high degree. In this respect convenience samples are always suspect.

Let us imagine that we pick two samples of 200 languages. Sample A represents 50–100 families from different continents in a rather even way (using, e.g., variety sampling), and sample B represents 190 Austronesian languages and 10 Tupian languages from South America. If the task of the typologist was to make inferences

[2] Note that multiple regression was not applied in typology before the 21st century, and it was only mixed modelling that paved the way for really developing statistical modelling approaches in the field. That is also the basis on which B&GN's proposal for methodological robustness builds. We need to contextualize earlier practices in language typology to the methodological discussion and development at the time.

about the world's languages based on these options, without a question sample A would represent the variation in the population better than the convenience sample B. No statistical method could remedy such deficiencies in a convenience sample. Obviously, this is an artificial example and we do not expect that any typologist would use such datasets in reality, but it shows where the sampling-does-not-matter thinking could lead if taken to the extreme.

Instead of rejecting sampling methods in favour of statistical modelling and regressing back to convenience sampling, we propose an alternative. Sampling is still needed to ensure that the sample represents variation in the population to a sufficient degree. Biases can then be assessed in a next step via statistical modelling. This issue should also be clearly understood when using data that the researcher has not collected themselves, as is very often the case in quantitative language typology which is increasingly moving towards a data science approach to cross-language diversity. By allowing larger samples variety sampling techniques represent the whole range of variety in the population much better and thus also serve as useful datasets for statistical modelling. As a generic principle we thus propose to marry variety sampling with statistical modelling to ensure representativity as well as bias control. It should also be remembered that variety sampling continues to be important for its original purpose, namely in exploratory work on phenomena whose cross-linguistic variation has not yet been charted.

4 The statistical model

The approach adopted by B&GN is to use exactly the same data as in the replicated studies. This is smart, as it allows them to assess the extent to which inferences depend on the methodological approach. To assess the robustness of results, B&GN then use a statistical model they have presented earlier (Guzmán Naranjo and Becker 2022). That method uses phylogenetic regression to control for genealogical relationships between languages. This means that the model builds an intercept for each language, but those intercepts are adjusted to be correlated with the structure of the tree. This offers a clear improvement to earlier approaches in which genealogical relations were modelled, for example, by building a random intercept for each language based on their affiliation to the highest taxa of a language family (e.g., Jaeger et al. 2011; Sinnemäki and Di Garbo 2018).

While phylogenetic regression offers flexible and powerful ways to control for phylogenetic biases in typological research, it raises some technical and conceptual questions. For example, once we incorporate the family tree into the regression model, we need to decide which tree to use or whether it even matters. B&GN do not explicitly address this question, although they state that the time depths of the trees are not needed, only an approximation of the distances between languages within a family. This may be understandable when phylogenetic relationships are treated as a confounding factor. However, this would seem to lead to a narrow interpretation of the results as mere synchronic patterns, not reflecting dynamic processes which have been of great interest in quantitative language typology over the last 20–25 years.

Besides controlling for phylogenetic bias B&GN's method assesses the effect of language contact with a Gaussian process. The Gaussian process is built on a measure of geographic distances between the language populations, computed from longitudes and latitudes for each language community. Again, this is an improvement compared to earlier research in which geographical proximity has been modelled by building a random intercept for each language based on the geographical area (e.g., macroarea) in which the majority of the language community resides. But using a Gaussian process for addressing contact phenomena makes some assumptions that are not explicitly addressed in B&GN. They are not unique to B&GN but underlie much work in quantitative language typology. The model assumes, for example, that the smaller the geographical distance is between language populations, the more likely there will be similarities between the languages due to language contact. Although this is a reasonable assumption and one that echoes much work in contact research, we do not yet know how well it holds empirically.

Languages used by neighbouring populations may certainly be similar due to their mutual social interactions. However, it is well-known that neighbouring language communities may have very strong language ideologies that prevent contact effects from spreading despite otherwise heavy social contact (e.g., Rodríguez-Ordóñez 2019). Moreover, strong language ideologies can encourage the development of innovations that make neighbouring languages more dissimilar to each other. Several examples have been recently discussed in the contact literature that indicate that increased dissimilarity in contact situations is much more prevalent than has been assumed thus far (e.g., Braunmuller et al. 2014; Evans 2019).

These issues are naturally not unknown to quantitative language typologists, but they nevertheless emphasize the need for discussing underlying assumptions in our models. Our point is the following: language contact is a social phenomenon, and while a novel geographical proxy for contact may be better than earlier geographical proxies, it is still a mere geographical proxy and leaves the fundamental issue unaddressed. Instead of reaching for better geographical proxies of language contact, we should assess language attitudes, language ideologies, and the degree of social interaction between populations. Doing so would allow us to assess the extent to which some measure of geographic distance might approximate the social contact dynamics between language populations. Ongoing work in sociolinguistic typology allows us to assess such hypotheses in the near future (e.g., Hartmann 2024; Kashima et al. 2025).

5 Conclusions

Overall, we welcome B&GN's call to take replication more seriously in language typology. The principles they propose are much needed and should be adhered to by practitioners in the field. In our commentary, we raised some issues related to robustness, transparency, sampling, and the statistical model B&GN use for assessing the results in the four case studies. While we endorse the idea of controlling bias via statistical modelling, our most critical comments of the target article concern the authors' views on sampling. We strongly disagree that convenience sampling would be sufficient and argued instead that future work should combine principles of variety sampling with statistical modelling. Their model appears promising, although it may assume a deep learning curve and raise issues related to career development and the division of labour within typology.

References

Aikhenvald, Alexandra Y. 2010. *Imperatives and commands*. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2017. Introduction: Linguistic typology – setting the scene. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge handbook of linguistic typology*, 1–36. Cambridge: Cambridge University Press.

Becker, Laura & Guzmán Naranjo Matías. 2025. Replication and methodological robustness in quantitative typology. *Linguistic Typology*. https://doi.org/10.1515/lingty-2023-0076

Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols*, 415–444. Amsterdam: John Benjamins.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heike Narrog (eds.), *The Oxford handbook of linguistic analysis*, 2nd edn., 901–923. Oxford: Oxford University Press.

Braunmüller, Kurt, Steffen Höder & Karoline Kühl (eds.). 2014. *Stability and divergence in language contact: Factors and mechanisms*. Amsterdam: John Benjamins.

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. *Linguistic Typology* 14(2–3). 253–286.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. The world atlas of language structures online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at https://wals.info).

Evans, Nicholas. 2019. Linguistic divergence under contact. In Michela Cennamo & Claudia Fabrizio (eds.), *Historical linguistics 2015: Selected papers from the 22nd international conference on historical linguistics, Naples, 27–31 July 2015*, 563–592. Amsterdam: John Benjamins.

Greenberg, Joseph H. 1978. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of human language I: Method and theory*, 61–92. Stanford: Stanford University Press.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.

Hartmann, Frederik. 2024. Geospatial models for analyzing contact effects. In *Symposium "Language Contact & Linguistic Adaptation", 30–31 May 2024*. Helsinki.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15(2). 281–319.

Kashima, Eri, Francesca Di Garbo, Ruth Singer & Olesya Khanina. 2025. The design principles of a sociolinguistic typological questionnaire for language contact research. *Language Dynamics and Change* 15(1). 1–103.

Keinänen, Satu. 2025. *Functional extensions of evidentials: A cross-linguistic study*. Helsinki: University of Helsinki PhD dissertation.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4. 307–333.

Miestamo, Matti. 2003. *Clausal negation: A typological study*. Helsinki: University of Helsinki PhD dissertation.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296.

Nichols, Johanna, Jonathan Barnes & David A. Peterson. 2006. The robust bell curve of morphological complexity. *Linguistic Typology* 10(1). 96–106.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13(2). 293–315.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method for language sampling. *Studies in Language* 17(1). 169–203.

Rodríguez-Ordóñez, Itxaso. 2019. The role of linguistic ideologies in language contact situations. *Language and Linguistics Compass* 13(10). e12351.

Sinnemäki, Kaius. 2011. *Language universals and linguistic complexity: Three case studies in core argument marking*. Helsinki: University of Helsinki PhD Thesis.

Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141.

Stassen, Leon. 1997. *Intransitive predication*. Oxford: Oxford University Press.

Stassen, Leon. 2009. *Predicative possession*. Oxford: Oxford University Press.

Tomlin, Russell. 1986. *Basic word order: Functional principles*. London: Croom Helm.

Widmann, Thomas & Peter Bakker. 2006. Does sampling matter? A test in replicability, concerning numerals. *Linguistic Typology* 10(1). 83–95.