Commentary

Laura Becker* and Matías Guzmán Naranjo

Authors' response to "Replication and methodological robustness in quantitative typology"

https://doi.org/10.1515/lingty-2025-0063
Received June 28, 2025; accepted June 28, 2025; published online July 28, 2025

1 Introduction

We want to start our replies by thanking all the authors who took the time to read our paper[1] and write very insightful commentaries. We also thank the editorial team of *Linguistic Typology* who suggested turning our contribution into a target paper, which made this inspiring discussion possible in the first place.

Although we do not agree with every point made in the commentaries, we are glad to see that there is a consensus that replication, robustness, openness and transparency should be a fundamental part of the field and that there are many aspects related to these issues that still deserve a closer look. The main objective of our paper was to draw attention to the need for more replication, robustness tests and transparency in quantitative typology, and given the general reactions, it seems to us that we all agree on that. We also realize that the methodological robustness issues discussed in our paper connect with many more, perhaps more important questions regarding replication not only in quantitative typology but typology and linguistics more broadly.[2] For reasons of space and our areas of

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

2 With that, our discussion ties in with other recent discussions around replication, reproducibility and robustness in corpus linguistics (e.g. Egbert et al. 2025; Flanagan 2025; Gries 2025; Laitinen and

***Corresponding author: Laura Becker [laura bɛkɐ]**, Department of General Linguistics, University of Freiburg, Freiburg im Breisgau, Germany, E-mail: laura.becker@linguistik.uni-freiburg.de. https://orcid.org/0000-0002-1835-9404

Matías Guzmán Naranjo [matias gusman naranxo], Department of General Linguistics, University of Freiburg, Freiburg im Breisgau, Germany, E-mail: mguzmann89@gmail.com. https://orcid.org/0000-0003-1136-6836

expertise, we could not address many of these related aspects in our study. We are happy to see that the commentaries drew these connections and offered diverse perspectives and approaches to replication, reproducibility and transparency in typology. We found a number of recurring themes throughout the commentaries, which we address together in Section 2. We then reply to the individual commentaries in Section 3.

2 Recurring themes in the commentaries

2.1 Methodological pluralism

It is worth reiterating that it is not the aim of our study to claim our methods are better than the rest. We do think that for certain types of question and data, some methods may be more suitable than others, but at this point, methods have not been compared sufficiently in (quantitative) typology to have a good answer in most cases. This was one of the reason for our study, i.e. to examine how different methods and their results compare, and to work towards more insights on methodological robustness in quantitative typology. We are happy to see many different techniques used in typology; naturally, we would not advance as a research field if all studies were to use the same method. We do believe, however, that if two techniques yield diverging results, we should think about why this happens, as it can teach us about the certainty we have in the results but also about the methods that we use. It is possible that the differences reflect a more fundamental problem with one or both methods, or they could result from conceptual issues in the study design (as pointed out by **Tallman** and **Mauri & Sansò**).

One point of criticism raised by **Di Garbo** is that we suggest in our paper that statistical techniques, like the ones we use, have a higher degree of reliability than traditional sampling approaches. We fundamentally agree that "not all typological investigations can, or should, be designed around statistical bias control." We do not claim that other, e.g. qualitative, approaches to typological research are inferior, or that our methods are the only correct ones. Moreover, we agree that the linguistic classification and annotation steps should take precedence over the type of analysis in quantitative studies. It is more important to get the linguistic facts right (as best as we can) than to use sophisticated statistical techniques.

Rautionaho 2025; Schweinberger and Haugh 2025a,b), phonetics (e.g. Coretta et al. 2023; Roettger 2019; Roettger et al. 2019) and linguistics more generally (e.g. Berez-Kroeker et al. 2018b; Grieve 2021; Sönning and Werner 2021; Vasishth and Gelman 2021).

Another point that relates to methodological pluralism is brought up by **Tallman**. He notes in footnote 2 that it may not be very surprising that we obtain partially different results in our robustness analysis of Shcherbakova et al. (2023), because phylogenetic relatedness is modeled differently in the two approaches. We agree that, of course, differences in the results could be due to the different types of models employed. But without concrete empirical testing, we simply do not know a priori to what extent the results necessarily differ or not. One of our goals was to contribute to a better understanding of how these two approaches compare in practice. To the best of our knowledge, this has not yet been tested in quantitative typology.

2.2 Transparency in typology

Several commentaries (**Gawne et al.**, **Haspelmath**, **Mauri & Sansò**, **Miestamo & Sinnemäki** and **Hammarström**) addressed the issue of transparency in the (quantitative) typological workflow. We take this as a sign for transparency at all steps of qualitative and quantitative typological work to be one of the most important aspects of replication and reproducibility, and that we can still do better in typology to achieve that. We are glad to see many different suggestions in the commentaries (which we will react to in our replies below); it is clear that we need this kind of discussion and a collaborative effort for better transparency standards in typology. Based on their previous work and expertise as co-chairs of the Linguistics Data Interest Group (LDIG), **Gawne et al.** highlight the importance of clear standards for data sharing and data citation for transparency but also reproducibility. We more than welcome the integration of the Austin Principles of Data Citation (Berez-Kroeker et al. 2018a) and the Tromsø Recommendations for Citation of Research Data in Linguistics (Andreassen et al. 2019) in typology. We also want to mention that *Linguistic Typology* has recently updated their data sharing policies, which, we hope, will contribute to an increase in transparency in typological work.[3]

Furthermore, we are glad to see that many authors of the commentaries agree with our suggestions for "Guidelines for better replicability in typology", and we more than welcome **Mauri & Sansò**'s recommendations for additions (cf. Section 3.2). We considered our guidelines to be a first proposal for transparency in typology, which obviously needs to be refined and complemented in a collaborative way. Hopefully this discussion can serve as a starting point for such an endeavor.

3 Cf. https://www.degruyterbrill.com/publication/journal_key/LITY/downloadAsset/LITY_Information_for_Authors.pdf and https://degruyter-live-craftcms-assets.s3.amazonaws.com/Policy-2_20240625.pdf.

2.3 Incentives in science

The commentaries by **Haspelmath** and **Gawne et al.** raise the issue of incentives in science in relation to openness, reproducibility, and replication. **Haspelmath** in particular notes that there are not many incentives for researchers to pursue replication and robustness research, and that it is more attractive to present new findings than to revisit old ones.[4]

Some incentives are institutional, and it is difficult for individual linguists to change them. The bias of most external funding towards projects that promise innovative research (at the level of the research questions as well as the methodology) can only be addressed if there is a change in the selection criteria of the funding organizations.

Other incentives, however, can be improved by linguists more directly. One of the major disincentives for replication and validation work is that such work is less likely to be published in the form of journal articles, which probably counts as the most important type of output in linguistics. Not all journals consider replication studies for publication, and most of the time, they are also expected to provide some innovative theoretical contribution (in addition to their methodological contribution). Linguists who are involved in the editorial work of journals can help to improve this situation by explicitly welcoming replication and validation studies, and/or by offering specific formats to include this type of work. *Linguistic Typology* has obviously done so in the case of our study, and by organizing this discussion, the editors have shown that replication, validation, robustness checks and transparency in typology are important questions that should be discussed in the community. *Linguistic Typology at the Crossroads* also accepted our proposal for a special issue on *Replication and data transparency in typology*, but not all journals share this openness. We could imagine that a special section or submission format for replication studies in addition to the standard formats such as research articles, reviews, etc., could help remedy this situation. Flexible formats such as "reports" and "methodological contributions", which many journals already offer (including *Linguistic Typology*) could explicitly list replication studies or robustness analyses as possible types of studies for this format. Another important incentive that we can influence is how these types of studies are viewed on our CVs by the linguistic community. Hiring committees might not rank replication studies as equally important or prestigious as theoretically innovative studies. Agreeing with **Dryer**'s commentary, we believe that carefully checking and evaluating previous findings and contributing to better methodology in our research field should be viewed as equally important as new findings.

4 Cf. also Roettger (2021: 1231–1236) for a discussion of incentivizing confirmation over exploration in experimental linguistics.

2.4 Statistical training and competence in the field

One point raised in the commentaries by **Di Garbo**, **Haspelmath** and **Miestamo & Sinnemäki** is that not all typologists can (or should be required to) use the techniques employed in our study and other similar quantitative statistical methods. It was not our intention with this paper to raise the issue of statistical training and knowledge in linguistics, and we may not be able to give an objective and qualified answer to this question, but we have some thoughts on the matter.

Miestamo & Sinnemäki comment that while our model appears promising, it has a steep learning curve. The authors also suggest that it may "raise issues related to career development and the division of labor within typology." We are not entirely sure what exactly they mean with their last statement concerning the division of labor, so we may either agree or slightly disagree with the authors. We ourselves are linguists by training (and not computational linguists or data scientists), and as far as we can tell, there is no fundamental barrier for PhD students in linguistics or typology to learn quantitative, statistical methods. While some degree of specialization and division of labor is an integral part of scientific work, in an ideal world, typologists working on quantitative projects should themselves have some understanding of the computational techniques. Otherwise, inconsistencies between the theoretical assumptions and the models used for analysis can easily arise, which is problematic, as **Tallman** and **Coupé** pointed out in their commentaries.

We agree with **Miestamo & Sinnemäki** in that quantitative methods for typology can have a steep learning curve, and we agree with **Di Garbo** in that qualitative approaches are equally important and valid and should by no means be viewed as inferior to quantitative approaches to typology. Of course, typologists who primarily carry out qualitative work do not necessarily need a strong statistical background. Thus, while not all of typology can and should be quantitative or computational, we do believe that typologists whose main focus is on quantitative approaches should strive to develop and improve their methods. The same way we expect (new) linguists in language documentation to be familiar with modern language documentation and fieldwork techniques, we do not see a good reason why (new) quantitative typologists should not be acquainted with modern quantitative techniques.

Related to statistical training, **Haspelmath** suggests that given the small size of the field, having training sessions for typologists is difficult to implement. Citing McElreath (2020), **Haspelmath** also writes that "it is not immediately clear what aspects of statistics one needs most urgently". We understand both concerns. We do not have a solution for how to train future typologists best, but in our experience, having statistical coursework is more and more common (though not necessarily

sufficient) in linguistics programs at both undergraduate and graduate levels, and we also see an increasing number of workshops at conferences or summer schools on statistics and programming skills for linguistics including typology. So at the very least, we do not think that statistical training in a small field like typology is generally impossible. In addition, by now there is a large number of introductions to statistical methods in linguistics and beyond that are useful for typologists at any career stage to learn about more fundamental as well as more advanced statistical methods (e.g. Bolstad and Curran 2017; Desagulier 2017; Gelman et al. 2013; Gries 2009, 2013; Levshina 2015; McElreath 2020; Winter 2020).

2.5 Terminology around replication, reproduction and robustness

Several authors, **Gawne et al.**, **Miestamo & Sinnemäki**, **Tallman**, **Shcherbakova et al.** and **Hammarström**, have commented on our terminological choices. As was noted by **Tallman** and **Shcherbakova et al.**, there are different ways to define replication, and we would argue that there is no consistent standard for terminology around this topic in linguistics or typology. After discussing among ourselves and with other typologists, we opted for "replication" because it seemed the most common and easily accessible label that many typologists would immediately recognize and understand without additional explanation. In doing so, we followed Hartmann (2022), who also uses the term "replication" throughout his study that applies new methods to old data. In addition, we chose to include "methodological robustness" in our title to signal that our study tested for the robustness of results across different statistical methods. Being aware of potential issues with our terminological choices, we tried to be as explicit as possible and therefore defined our use of the notions of robustness, replication and replicability in Section 2.1 of our study.

However, we do not have strong opinions regarding our terminological choices and are happy to go along with what the linguistic and typological community prefers. Several of the commentaries make concrete suggestions as to what our study does instead of replication:

- **Tallman**: tests of robustness
- **Hammarström**: test the robustness
- **Shcherbakova et al.**: robustness analysis
- **Gawne et al.**: reproduction

"Robustness tests" or "robustness analysis" seem to be preferred labels, which we have adopted in our response here. On a more general note, we find it interesting that there is no clear agreement between all authors about what the optimal terminological

choice is. This shows that it is not always trivial to transfer the use of concepts such as replication, reproduction, or robustness tests from other (linguistic) disciplines to typology without a discussion in the community like the present one.

2.6 Modeling spatial relations between languages

Another main theme in the commentaries is that of how we model space and contact between languages. For instance, **Dryer** comments on the fact that distance is relative, and that geographic features can have an impact on contact, and **Shcherbakova et al.** and **Hammarström** note that our distance metric is only an approximation. **Miestamo & Sinnemäki** mention more complex contact scenarios, where two languages can, in fact, diverge from each other due to contact. **Coupé** discusses that even within Gaussian processes, there are several options for a covariance kernel, which could produce different results with identical distance matrices. We agree; modeling spatial relations between languages is difficult, and we are well aware of numerous shortcomings of our method. At the same time, we do not think that we should simply ignore spatial relations between languages in statistical modeling of typological questions, and we are trying to work with techniques that are appropriate modeling choices to the best of our knowledge at this moment in time. It is not our intention to argue that our specific take on spatial modeling is a final or optimal solution, or that Gaussian processes can capture all complexities found in language contact. There are many things we do not capture, some of which are brought up in the commentaries, some of which are not.

A first general point we want to make is that we deliberately chose to work with the *brms* package in R to make our methods more accessible to the wider typological audience. Writing our models in Stan would have allowed for more flexibility. While we are proficient in Stan and have proposed more precise techniques in the past (Guzmán Naranjo et al. 2025; Guzmán Naranjo and Mertner 2023; Urban and Guzmán Naranjo 2025), we think that the methods used in our study represent a reasonable compromise between modeling complexity, model adequacy and accessibility. Of course, others may argue that the optimal trade-off is more on the side of modeling complexity and model adequacy.

2.6.1 Modeling particular contact scenarios

From the qualitative literature, we know that many types of contact effects are not uniform across the world, and some can be relevant for particular regions only or even affect single pairs of languages. Related to this, **Miestamo & Sinnemäki** highlight the empirical observation that, depending on language ideologies, contact

can also lead to linguistic differentiation instead of convergence. Another example is mentioned by **Dryer**, who notes that marriage patterns in Papua New Guinea lead to women "marry[ing] into villages where the language is more than ten languages away from their native languages". In such a setting, contact between two languages thus leapfrogs several languages. As far as we can gauge, particular and potentially idiosyncratic effects of language contact like these ones are the hardest contact-induced effects to model computationally. Building a general model structure that can take into account language ideologies is extremely difficult and moreover requires very detailed socio-linguistic data.[5] We can say for certain that Gaussian processes are likely the wrong tool for this type of linguistic situation, but we do not know what would be an appropriate tool. We remain excited about new developments in this area.

2.6.2 Euclidean versus other types of distances

In their commentary, **Shcherbakova et al.** criticize our approach for using Euclidean distances. They argue that Euclidean distances are only (poor) approximations of the real separation between communities because they assume a flat and deformed earth. Regarding Euclidean distances versus other types of distances, Guzmán Naranjo and Jäger (2023) introduce and compare topographic and walking distances and show that, very often (although not always), topographic and walking distances do indeed perform better than Euclidean and Geodesic distances. These types of distance metrics can be used in Gaussian process models, but they require the model to be written in Stan. As was highlighted by **Miestamo & Sinnemäki**, **Haspelmath** and **Di Garbo**, the modeling choices presented here may already present a steep learning curve. We agree with that, and we wanted to encourage others to use our models and therefore avoided writing in Stan directly, as it requires familiarity with another programming language in addition to R.

Furthermore, there are some additional, but minor complications when using more realistic distance measures. Topographic and walking distances require some additional work to get them into shape to be included in a Gaussian process. Computationally these distances are also much more costly to compute, and thus, they will necessarily represent an additional layer of complexity for researchers.

---

5 Socio-linguistic data seem to be difficult to gather cross-linguistically. **Mauri & Sansò** mention another group of socio-linguistic variables pertaining to literacy, and suggest crowd-sourcing and community-based data collection as a possible way to improve our current situation. We fully agree; having more fine-grained, cross-linguistic data about socio-linguistic variables is a necessary step towards including these properties in statistical models.

Another relevant point that we do not discuss explicitly in the paper is that exact Gaussian processes are computationally very costly. This means that they scale poorly with the number of languages in the sample. While fitting a Gaussian process with 100 observations is very fast, it becomes slower once a sample has more than 500, and anything above 1,000 observations will be very slow. A practical solution is the use of approximate Gaussian processes (Riutort-Mayol et al. 2023), but these are only available with Euclidean distances. The consequence is that if one wants to work with very large datasets like Grambank, exact Gaussian processes with more complex distance metrics become impractical to work with.

2.6.3 Macro-areal effects versus local contact effects

Both **Dryer** and **Seržant** note the apparent contradiction between the assumptions of a Gaussian process that contact quickly decays to 0 after a few hundred kilometers, and the potential existence of large macro-areal effects in the typological distribution of linguistic features. We agree with Dryer's interpretation of the situation, namely that large contact areas emerge as the result of serial, (relatively) short-distance contact events. What he calls snowball effects, Guzmán Naranjo et al. (2024) have also referred to as water bucket effects.[6] The point is that a series of local contact events can produce very large linguistic areas. This situation does not represent an issue for Gaussian processes, and water bucket effects can be captured without problem: Gaussian processes can build large contact zones from local contact situations.

2.6.4 Modeling spatial relations versus real contact relations

Both **Hammarström** and **Seržant** make an important conceptual distinction between the spatial effects that we model and the (many different types of) contact situations that have occurred and occur in the real world. This leads to the important overall question which subsumes most other comments discussed in this section, namely how the spatial relations that we model actually relate to real contact phenomena between languages.[7] In short, we do not model language contact directly, but use spatial relations between languages as a proxy. It is extremely difficult, if not impossible at this point, to build a statistical model that captures (direct) language contact on a global scale in an adequate way. We have pointed out some of these complexities in our replies above, but we feel that this fact is still under-appreciated

6 The idea is the image of a human chain transporting water buckets to fight a fire. Each person only moves the bucket a short distance, but the bucket can travel a long distance.

7 In a way, this question is an example of what **Tallman** and **Coupé** argue for, namely careful (or better) integration of theoretical assumptions in the statistical analysis.

in the typological community, and that some linguists (not the authors of this discussion) right out dismiss approaches like ours because certain details about language contact are not properly captured. We sympathize with the frustration from an expert's perspective on language contact, but we do not think that the consequence should be to stop trying to capture contact in quantitative models, even if it is via spatial relations. What can help to advance modeling techniques of contact effects in quantitative typology is to start out with smaller areas and more controllable (socio-)linguistic realities, where we can try to find modeling solutions that fit the reality better. This is what we have tried to do, for instance, for asymmetric language contact and expansion effects for Polynesian languages in Guzmán Naranjo et al. (2025) and for the Americas in Urban and Guzmán Naranjo (2025).

To sum up, we fully agree that our approach to space in the present paper cannot capture but a fraction of the complexities of real language contact. At the same time, we believe that the methods we employ in this paper are a reasonable compromise for most studies, and that they offer an improvement over other quantitative approaches that include no spatial component at all. We will continue to work on developing and testing new techniques that take the complexities of real language contact situations more seriously, but it remains challenging to adapt and use such techniques in large-scale typological studies.

3 Responses to the individual commentaries

3.1 Francesca Di Garbo

3.1.1 Methodological pluralism

In her commentary, **Di Garbo** provides a wider perspective to our study by distinguishing four types of researcher profiles in typology, from documentary linguists to statisticians. She shows how different researcher profiles and perspectives lead to different types of typological studies that make up an incremental workflow from primary data collection to exploratory comparative work to statistical modeling of distributions in space and time. **Di Garbo** highlights that each of these perspectives has and requires different methodologies, and she rightly points out that "not all typological investigations are amenable to statistical modeling, it is important to underscore that inferential statistics may not be the only way to validate research results in our field". We fully agree with this statement; in our paper, we focused on quantitative typological studies only, for which we do think that statistical modeling is relevant and important. For other types of typological research, especially qualitative and more exploratory work, statistical modeling

and statistical bias control are irrelevant. **Di Garbo**'s commentary thus emphasizes the importance of the plurality of approaches and perspectives in typology and makes a call for an inclusive view of the field, where none of the approaches "should be conceived of as scoping above the others." We completely agree with this point. In particular, we second the importance of descriptive and qualitative exploratory linguistic work as particularly relevant. Without a solid understanding of the linguistic realities, quantitative and statistical approaches cannot contribute much to a better understanding of why linguistic properties are distributed the way they are.

3.1.2 Sampling and statistical testing versus statistical bias control

Another critical point raised by **Di Garbo** is that sampling plus statistical testing is a useful method to make hypotheses and discover new potential paths of research. She writes that "[s]tatistical testing can thus be a valid way to (start) investigating small to moderately large typological datasets." In principle, we agree with this point in that we do not want to promote methodological gatekeeping for sample-based typological studies and suggest that they should only be carried out and published using advanced statistical modeling. As mentioned above, we also agree that much of typological work does not need to be quantitative at all.

What we would argue, though, is that if a primarily qualitative typological study is supposed to include a first quantitative overview of the cross-linguistic distributions, showing the raw counts and proportions is more useful than additional statistical tests. As we mentioned in Section 7.2.2 of our paper, others have shown in detail why statistical tests such as Chi-Square tests or *t*-tests are often not very useful. Their assumptions are not met in most cases, which makes their results irrelevant. Therefore, in cases of primarily qualitative studies, we see no issues with simply reporting the raw distributions, which can also serve to formulate hypotheses to be tested in a further study.

In cases of typological studies with a clear quantitative focus, we are less certain about how suitable some form of probabilistic sampling with statistical tests may be, and we think that there is evidence for statistical modeling with statistical bias control to be advantageous. We base this on several previous studies that have shown how clear effects found in studies using no or less detailed statistical bias controls are considerably weakened or disappear once such controls are included (e.g. Becker et al. 2023; Guzmán Naranjo et al. 2025; Hartmann 2022; Hartmann et al. 2024; Jaeger et al. 2011; Roberts et al. 2015; Van Tuyl and Pereltsvaig 2012). Not all of these studies were based on probability samples to begin with, but some were, which has two possible consequences. We can either use statistical bias controls to model the remaining dependencies between the languages in the sample, which often

represent contact or areal relations between languages. The alternative is to incorporate the contact and spatial component more seriously in the sampling method used, where the focus has traditionally been more on phylogenetic relations between languages. Dryer's (2018) study that we tested (and much of his previous work using macro-areas) is a case in point.

3.2 Caterina Mauri and Andrea Sansò

The commentary by **Mauri & Sansò** addresses the question of replicability and transparency in typology further upstream, i.e. at the stage of data collection and data annotation, which they illustrate with a case study on the effect of literacy on grammatical structures. We did not focus on this level of replication and replicability in our study and could not discuss it much for reasons of space. We agree with **Mauri & Sansò** that it is a very important, probably even more important aspect of replicability in typology, since the data and annotation correspond to the foundation of any qualitative or quantitative typological study. If the data collection and annotation processes are not transparent, they are likely to create to more transparency and replicability issues further down the line.

We also fully agree with **Mauri & Sansò**'s two recommendations for typological datasets. First, they suggest that datasets should allow for the option of reversing binary or other, simplified classifications into earlier, more fine-grained distinctions whenever applicable. Their second recommendation is that comment fields should be used to document additional clarifications for individual annotation choices as needed. The third recommendation that the authors make for the data annotation process more broadly is to include what Cysouw (2007) introduces as a social layer. As we understand **Mauri & Sansò**'s and Cysouw's (2007) suggestion, this layer can go beyond an additional comment column and could be understood as a separate appendix or even as the primary version of the database, to which the final, simplified, curated and automatically processable dataset is added. This social annotation layer would then include a more general documentation of the research questions leading to the overall classification and annotation choices with relevant details of the classification and annotation process itself. It could also include more fine-grained information pertaining to single observations and annotations such as precise references, examples, and other comments about the authors' interpretation of the data. We have little to add other than to say that we fully endorse these recommendations and thank **Mauri & Sansò** for highlighting these important aspects of transparency in data collection and annotation in typology.

3.3 Martin Haspelmath

In his commentary, **Haspelmath** discusses six important aspects in relation to robustness and explains how they affect typological work. We agree with his first point on robustness across different languages samples, and we already addressed his point on statistical training in Section 2.4 and on incentives in Section 2.3. We respond to the other points below.

3.3.1 Alternative analyses and annotations

As we understand it, the issue of robustness in the annotation and analysis of linguistic data across studies is one of the most important aspects for **Haspelmath** in terms of replicability in typology. He writes: "This is perhaps the greatest stumbling block for reproducibility in comparative linguistics, because the comparative concepts used for cross-linguistic comparison are not standardized in the field, and traditional terms are often understood and used in diverse ways across scholars and subcommunities." We agree with his assessment of the status quo in that concepts and terminology are used in many different ways across studies, scholars and research areas. Yet, we do not think that this is an insurmountable issue; as suggested by **Mauri & Sansò**, it requires more transparent documentation of the annotation process and more explicit definitions of the concepts and terms used. We understand this point as a reminder to us all to be as explicit as possible in our work when defining linguistic terms, even if (or especially if) the terminology seems well established. We also think that including linguistic examples to illustrate each of the annotation decisions can help to remedy unclear terminology, as it allows the reader to check and follow the annotation decisions for themselves.

As for the robustness of findings based on alternative analyses and annotations, we think that if two different operationalizations of a comparative concept produce very similar results, it provides evidence for the general classifications to be linguistically meaningful. If the two operationalizations lead to different results, we should try to understand which of the theoretical decisions could have led to these differences.

3.3.2 Protection against cognitive biases

Haspelmath rightly points out that cognitive biases can have an impact on the results of a study. This aspect of robustness relates to what **Smith** discusses in his commentary as researcher flexibility or researcher degrees of freedom. **Haspelmath** notes that while in experimental research, blinding is a possible shield

against this type of bias, it is less so in comparative grammar studies. He concludes that "[t]his particular problem may not be so acute with large-scale studies of the type discussed by B&GN, but otherwise it is unfortunately quite typical of the field of (non-quantitative) linguistics." We partially disagree with this last point in that we think that cognitive biases are also highly relevant for large-scale quantitative typological studies as the ones that we focused on in our paper. There is a garden of forking paths when a researcher builds a statistical model, and every decision has an effect on the results of the model. Some of the many choices include: the statistical software used, choice of likelihood, pre-processing of predictors, linear versus non-linear effects, types and number of controls, multiple models versus multiple regression, Bayesian versus frequentist framework, prior selection, model checks, etc. Every single one of these choices has an impact on the final results of the model, and we do not have an agreed upon setup in typology (and likely never will). Researcher bias could easily influence these choices to guide the model toward the desired result. The degree of such biases, especially concerning modeling choices, is exactly what we wanted to assess in our study by testing previous studies for robustness using the original data but different statistical methods. As we mentioned before, methodological pluralism is desirable, and paired with replication studies of various kinds, it can help us to understand better the impact of cognitive biases in (quantitative) typology.

3.3.3 Encouraging team science

Regarding team science, **Haspelmath** mentions the "Many Labs" project, and discusses whether something similar could be implemented in typology "[…] where dozens of linguists working in different locations collaborate to improve the methodology of comparative grammar". This is a very interesting and innovative take on robustness testing and it would make much sense to combine efforts this way.[8] We see different ways how a typological "Many Labs" project could play out. On the one hand, it would be fairly easy to do something like a shared task in quantitative typology. One would need to provide a typological dataset and have several teams build different models to analyze the same data. On the other hand, we think that some existing typological work sort of fits **Haspelmath**'s "Many Languages" description, e.g. collaborative projects with a common theoretical framework such as the Leipzig Valency Classes Project (Malchukov and Comrie 2015a, b) or the Mainz Grammaticalization Project (Bisang and Malchukov 2020a, b). What we take away from this for the future is that a potential "Many Languages" collaboration is something to keep in mind and certainly worth a try to organize.

8 We should mention the "Many Speech Analyses" project in phonetics (Coretta et al. 2023) as a first implementation of the Many Labs approach in linguistics.

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

The commentary by **Gawne et al.** provides an insightful overview of their previous work (in part as the Linguistics Data Interest Group) that has tried to evaluate the status quo in linguistics with regards to transparency, openness and reproducibility. For typology, the authors report that the field still has a long way to go, and that most studies do not meet minimal standards for transparency and openness, which means that the requirements for reproducibility and replicability are still often not met. As mentioned in our general reply in Section 2.2, we whole-heartedly agree with all points raised in this commentary. We strongly support the type of work such as the Austin Principles that **Gawne et al.** are doing for data transparency to become a standard in typology and in linguistics more generally, and we thank the authors for their detailed commentary and expert insights.

3.5 Kenny Smith

3.5.1 Researcher degrees of freedom

Smith correctly points out that our robustness analyses relate to a well-known phenomenon from experimental psychology, namely that researchers are faced with a series of choices when analyzing data, and that these choices can have a large impact on the results of the analysis (Gelman and Loken 2014). Two researchers analyzing the same dataset can reach very different conclusions if they make slightly different (conscious but also unconscious) choices, each of which can lead to more analytical differences further down the line. The difficulty lies in the fact that it is often not obvious that there is a single set of correct choices. We appreciate the wider perspective from experimental psychology that **Smith** offers, and the reminder that researcher degrees of freedom are well known in other disciplines. Our impression is, though, that this has not been discussed much in general linguistics, one notable exception being the "Many Speech Analyses" project, which examined analytical flexibility in phonetics (Coretta et al. 2023). **Smith** is correct in pointing out that, against this background, it is not surprising that we can replicate some but not all results in our study despite using the original data.

While experimental fields have found pre-registration to be a useful mechanism to protect against researcher bias, it is less clear that pre-registration would be useful for more methodological robustness in typological research. In qualitative or more exploratory typological work, many of the decisions regarding how to categorize the

data and which phenomena to include and to exclude are made and updated after having seen and analyzed the linguistic data. To an extent, this also holds for quantitative typological studies that build their own sample with grammars.

3.5.2 Triangulation with experimental linguistics

Smith brings up another important aspect, i.e. the relation between typology and experimental linguistics, the latter of which can help "to test cognitive and interactional mechanisms hypothesized to be responsible for potential universals." This was not a main point in our paper, also because we have only limited experience with experimental approaches, but we welcome the insights from this broader and interdisciplinary perspective.

Smith suggests that experimental data can help us better understand the causal mechanisms behind typological generalizations, something observational typological studies cannot do. We generally agree that some research setups are more adequate for investigating certain types of questions, and a division of labor, or triangulation, makes sense from this perspective. The difficulty emerges, again, with cases of disagreeing results between experimental and typological studies. **Smith** provides two very insightful examples of such cases. We will react to the first example, as it concerns a topic that we also explored in previous work, namely the relation between sociolinguistic factors and linguistic complexity (cf. Becker et al. 2023; Guzmán Naranjo et al. 2025). In both cases, we failed to find clear, convincing evidence for sociolinguistic correlates of linguistic complexity. In contrast, Smith (2024) reports on an artificial language learning experiment that supports the presence of mechanisms proposed in the typological literature to account for an association between sociolinguistic factors and linguistic complexity. In such a situation, the important question arises: how can we understand the discrepancy between the results? Smith mentions two hypotheses: (i) the factors identified in the experiments are outweighed by other factors in the wild, and (ii) natural language data cannot show the correlation with sufficient confidence. We agree, and we can think of a number of other potential explanations that can lead to the situation of finding an effect of, e.g., socio-linguistic factors on linguistic complexity in experimental studies but not in typological ones. We think that all these issues should be explored and subsequently discarded in order to understand diverging results:

- experimental studies:
  - the experimental design may not be suitable
  - the experimental study may not reflect natural language learning
  - the data analysis of the experimental study may have issues

- typological studies:
  - the study may not operationalize the actual socio-linguistic hypotheses well
  - the data collection and annotation may contain too many mistakes
  - the language sample may be too small to detect the (potentially weak) effects
  - the language sample may be wrong in just the right way, hiding the effects
  - the data analysis of the typological study may have issues

These issues all highlight the possibility that either the experimental or typological studies could lead to fundamentally incorrect results. This goes back to our main point: we can only increase our confidence about our findings with more transparency about the work process, with robustness tests and with replication. If at some point we reach high confidence about results from both experimental and typological studies, and these still diverge, we can then start to think about how and why they diverge. Currently, we do not believe that we can have high certainty about our typological results regarding sociolinguistic effects on linguistic complexity to begin with. Therefore, we should be cautious when trying to interpret differences between the typological and experimental results.

Finally, we agree with the conclusion of Smith who suggests to "view typology as part of a joint endeavor to uncover mechanisms shaping languages, where other methods (specifically, controlled experiments) are anyway required to test the critical mechanisms held to be responsible for candidate linguistic universals."

3.6 Matti Miestamo and Kaius Sinnemäki

3.6.1 Robustness across alternative linguistic analyses and annotations

Amongst other things, we argue in our paper that robust typological results should replicate across different datasets, different statistical methods, and different linguistic analyses and annotations. We gave Nichols et al. (2006) as an example for replication across different linguistic analyses. **Miestamo & Sinnemäki** point out that Nichols et al. (2006) is not a clear example of replication across different analyses, but rather "testing a theoretical claim through the lens of different linguistic phenomena using different types of data". As another example we can think of the classical topic of word order typology. Traditional approaches to word order universals (e.g. Dryer 1992, 2013a, b, Greenberg 1963, Hawkins 1983) operationalize dominant word order as the most common word order in simple, declarative sentences. Put differently, word order is represented in a categorical way, each language being assigned a single, dominant or basic word order (although some approaches allow for the value "no dominant order"). More recent, corpus-based

approaches to word order typology have argued for a gradient representation of word order preferences within languages (e.g. Futrell et al. 2015; Gerdes et al. 2019, 2021; Guzmán Naranjo and Becker 2018; Levshina 2019; Levshina et al. 2023; Östling 2015; Talamo and Verkerk 2022). Yet another recent take on word order typology, namely Grambank (Skirgård et al. 2023), uses a third strategy to operationalize word order preferences. Word orders are represented in a categorical way, e.g. as in Feature GB131 (*Is a pragmatically unmarked constituent order verb-initial for transitive clauses?*). However, different orders are split up into different features, which allows for single languages to be annotated as having more than one unmarked word order.[9] If we find that a cross-linguistic word order preference holds across all three types of operationalizations, our certainty about the underlying fact should be much higher than if this preference is only detected using one type of operationalization.

3.6.2 Full data transparency

We appreciate that **Miestamo & Sinnemäki** highlight in more detail than we did in our study that also qualitative, exploratory typological studies based on a variety or convenience sample should provide a comprehensive list of the languages as an appendix. At the very least, it would be useful to include the language name, its glottocode and the reference(s) used in such a list.

Furthermore, **Miestamo & Sinnemäki** are absolutely correct in noting that we missed important typological work that includes exhaustive examples for their annotation decisions; we should have been more careful in our phrasing. We thank the authors for bringing to our attention Stassen (1997, 2009), Miestamo (2003, 2005), as well several WALS chapters. We are probably still missing other work that should be mentioned here. Our main argument remains, though, namely that many sample-based typological studies do not provide such an exhaustive list of examples. Doing so may not be feasible in all cases, but we should strive to provide all relevant examples whenever possible to contribute to more transparency in typological work.

3.6.3 Sampling in quantitative typology

Miestamo & Sinnemäki's main point of criticism concerns the role that sampling should play in (quantitative) typological studies. We admit that we could have

9 For instance, Nez Perce is annotated as "1" (i.e. "yes") for Feature GB131, GB132 (*Is a pragmatically unmarked constituent order verb-medial for transitive clauses?*) and GB133 (*Is a pragmatically unmarked constituent order verb-final for transitive clauses?*), reflecting that all three orders are unmarked according to the grammatical description.

presented a clearer and less strong perspective in our study, and we thank the authors for addressing this issue in detail, giving us the opportunity to clarify our position.

Miestamo & Sinnemäki distinguish between probability samples, variety samples and convenience samples, the latter of which effectively corresponds to our proposal of using no manual sampling method. They argue that also in quantitative typology, variety sampling is a more sound approach than convenience sampling, and they show why this is the case discussing a theoretical example: if we systematically leave out phylogenetic or areal groupings of languages, statistical bias control cannot help to make up for missing representations of cross-linguistic variety. We fully agree with this argument and with **Miestamo & Sinnemäki**'s position, and we should have expressed our position in the paper more clearly and more carefully. We only referred to probability sampling in our study, and did not want to make any statement about variety sampling. We fully agree that variety sampling is advantageous over convenience sampling in quantitative typology, since statistical techniques cannot fix fundamental flaws that occur during the data gathering stage such as missing variety in heavily skewed samples. **Miestamo & Sinnemäki** make a strong argument for variety sampling when building linguistic datasets, as they have in the past (Miestamo 2005; Miestamo et al. 2016). In fact, in Guzmán Naranjo and Becker (2022: 605) we come to a very similar conclusion: "[…] we also show that strict probability sampling is not required with the statistical controls that we propose, as long as the sample is a variety sample large enough to cover different areas and families."

3.6.4 Modeling decisions

Miestamo & Sinnemäki address two issues regarding our model: the tree used for the phylogenetic term and the underlying assumptions of how space affects contact. We already discussed the latter comment in Section 2.6 and focus on the first comment here. The authors are correct in pointing out that including a phylogenetic tree in the model comes with additional conceptual and technical questions, and we cannot address them fully in this reply. What we will say is the following. The implications of using different trees in quantitative modeling in typology are not yet well understood. There are several ways of inducing tree branch lengths (e.g. Bouckaert et al. 2012; Guzmán Naranjo and Becker 2022; Wichmann and Rama 2021), but we do not yet know how much such differences in tree structures affect model results. Preliminary tests (not yet published) suggest that in practice, the impact is minimal to negligible because phylogenetic terms are flexible enough that they can accommodate variations in relative branch lengths. Of course, this should ideally be tested properly with data across different types of linguistic phenomena. Our second

answer is that whether we use branch lengths in thousands of years (based on some sort of imputation) or relative branch lengths in terms of number of shared nodes (what we do) cannot have an impact, provided that the relative relations are the same. What the model uses is the covariance between leaves. This covariance is calculated as Brownian motion based on the distance matrix.

The choice of the phylogenetic tree has no consequence for the interpretation of the results as synchronic or diachronic. The phylogenetic term is inherently diachronic, just as the Gaussian process term built on geographic distance is inherently spatial. The phylogenetic process captures the feature evolution along the tree. The intercepts of the model are then equivalent to a sort of steady state that expresses universal tendencies.

To conclude, these are valid concerns and questions, and we are currently preparing a study that goes into more detail regarding the conceptual basis for different technical ways of implementing phylogeny in quantitative typology.

3.7 Adam Tallman

3.7.1 Weak versus strong theories

Tallman proposes to focus more on the relationship between the linguistic theories and our statistical models. Following Fried (2020), he calls for a distinction between strong and weak theories. Strong theories are explicit about underlying assumptions and the causal relations to be tested, while weak theories are not. **Tallman** relates this distinction to the broader topic of replicability and our study by arguing that "the effects that lack robustness might relate as much to the fact that B&GN are attempting to assess weak theories." We agree that this is a very important issue not only regarding the replicability of (typological) studies but for a transparent scientific workflow in our studies more generally.

In our study, we wanted to highlight how complex the modeling decisions are when analyzing typological data and how much the results can depend on these decisions even when the data used is identical. We fully agree, though, that this only addresses one puzzle piece of robustness and replicability in typology, and that the question of how strong the theoretical underpinnings of our statistical analysis really are is in fact an equally if not more important question. We are therefore thankful that this aspect was brought up by **Tallman**, and we fully agree with his conclusion that statistical modeling can help us to formulate stronger hypotheses by being explicit about our assumptions and by making them testable.

3.7.2 Implicit assumptions

Related to the previous issue of strong versus weak theories that build the conceptual foundation for the statistical analysis, **Tallman** discusses implicit assumptions in Seržant (2021) and Shcherbakova et al. (2023), as well as in our robustness tests.

Regarding Seržant (2021), we do not fully agree. **Tallman** claims that Seržant has a (latent) theory about how the particular contact situation of Slavic languages has led to its skewed positioning in an East-West cline of person-number index decay, and that we could not have tested this theory properly with the type of data used. The goal of our robustness test was different, though. We primarily wanted to test the hypothesis about the East-West cline, i.e. whether we also find evidence for this cline when using statistical modeling. We tested for the East-West cline by explicitly comparing a model with a contact component versus a model with a contact and a cline component. It is a secondary question how such a cline could have originated, and what the concrete mechanisms were, that resulted in Slavic showing less paradigm decay than Germanic or Romance, and we agree with **Tallman** that this is something that we could not test with the type of data at hand.

As for Shcherbakova et al. (2023) and our robustness analysis, we fully agree with **Tallman** that both studies fell short on spelling out "explicit reference to causal theories about the diachronic relationship between grammatical coding" and that "the analysis could (or should) be regarded as mostly exploratory." Our operationalization of the replication tried to follow, as closely as we could, the hypothesis that the original study aimed to test, namely "[t]o investigate correlations between the amounts of coding on nominal words and verbs" and to "test whether a change in one domain implies a change in another domain" (Shcherbakova et al. 2023: 157). We agree that a stronger hypothesis or theory could be linguistically more informative and should take into account more complex relations between the linguistic features analyzed. This would also result in a more complex model structure that our setup could, in principle, be extended to.

3.8 Harald Hammarström

3.8.1 Better and poorer models

Hammarström correctly points out that there are two perspectives to our study: (i) robustness testing, which only makes sense if multiple methods are equally suitable, and (ii) trying to understand which models are better. We agree that we know for some techniques that they are better or poorer for modeling typological questions, e.g. we know that multiple regression is fundamentally better than fitting

many disjoint regression models. However, we do not think that we have a complete understanding of which methods are objectively better in all situations, also because it depends on the phenomenon and data at hand (as **Hammarström** also notes). To give a concrete example, we have argued before that phylogenetic effects and Gaussian processes are better than having effects by family and macro area (Guzmán Naranjo and Becker 2022). However, these results are only partial. In ongoing work, we find that phylogenetic regression can be suboptimal and perform considerably worse than family effects under certain conditions. This happens, for example, when a sample contains many isolates and small families. Similarly, while Gaussian processes are incredibly powerful, approximate Gaussian processes or splines can be more appropriate for large datasets because sampling exact Gaussian processes becomes exceedingly difficult with more than 1,000 observations.

For these reasons, in practice, we take (and need to take) both perspectives at the same time, i.e. checking for methodological robustness but also trying to understand which methods may be better for a given study. If the results are robust across methods, we learn that both methods are likely to work comparably well for the given data and question. If the results are not robust across methods, then we need to try to understand where the differences come from (as was also noted by **Dryer**). This can help to learn which method may be better for capturing a given linguistic phenomenon.

3.8.2 Standards for replicability

Hammarström claims that our suggestions for more transparency and better replicability are not ideal because we call for full code sharing. His argument is that code is neither sufficient nor needed for replication, and that in other disciplines like computer science, the standard is to share pseudocode specifications of the algorithms used. Pseudocode, as the name suggests, is a description of the steps in an algorithm without being executable code as such. There is no single convention for pseudocode, but the idea is that it is easily understandable without additional explanation for scientists who are familiar with programming, while being independent of a particular programming language. This may seem like a useful alternative to sharing code in a particular language, likely making it more sustainable in the long run (e.g. stable over decades and hundreds of years). Despite these potential advantages, we think there are several strong arguments against using pseudocode in linguistics.[10] Converting math and pseudocode into working code is not straightforward, as there is no single standard that can quickly be adopted, and it is yet another layer of the quantitative workflow that linguists would need to familiarize

10 We are not computer scientists and cannot comment on the standards used in that field.

themselves with. It is also an additional step where mistakes can happen that are likely to go undetected by the original author(s) because pseudocode cannot be executed and checked (Lamport 2009: 25). The fact that it cannot be checked or run is equally an issue for review, replicability/reproducibility and re-useability. Linguists may want to adopt methods from a previous study; they would then have to convert pseudocode back into actual code, which would likely lead to questions and additional choices that could be avoided without the conversion of code into pseudocode in the first place. Also, sharing pseudocode instead of the code used seems to add another skill to the (already long) list of skills needed to do quantitative typology.

There is an argument to be made about sharing model specification in statistical notation (we have done this in the past when we introduced new models, e.g. Guzmán Naranjo and Mertner 2023). While this is certainly important from a theoretical perspective, it may not be very useful for many linguists who would need to re-implement it to use the same model. Code, on the other hand, can be re-used and adapted easily for new projects.

If the concern is about whether code will run in the future, there are some solutions for that. The most robust one, but also the most costly in terms of time and effort, is to use docker (Merkel 2014) or similar containers.[11] Containers guarantee that the environment, from the operating system to the linear algebra libraries and to packages are the same every time. A less robust but still very robust alternative is to use a versioning software like renv or conda.[12] Conda environments are the standard in many fields. We provide both renv and conda environments for our code. The least robust but still useful solution is to simply share the session information (`sessionInfo()` in R or `session_info.show()` in python). For these various reasons, we disagree with **Hammarström** that pseudocode is necessary and sufficient, and argue that sharing the actual code is more helpful and transparent, at least for the foreseeable future in quantitative typology.

3.8.3 Some details on our models and method

Hammarström makes two statements about our model and methods that are not entirely correct and that we want to clarify here. He writes: "However, as employed in the replication study, the same hyperparameters apply across the entire sample, which means that the influence decays the same way per kilometer (or, actually, per lat/lon degree) over the whole world." This is not completely correct, at least not for all studies. We use grouped Gaussian processes by macro-area for the robustness test of Dryer (2018), which means that the parameter is shared by all languages in the

11 Cf. https://www.docker.com/resources/what-container/.

12 Cf. https://rstudio.github.io/renv/index.html and https://docs.conda.io/en/latest/.

same macro-area, but it allows for variation across areas. In our experience, however, this does not matter much in practice in most situations because Gaussian processes are very flexible and can adapt well to different areas even if one builds one single Gaussian process. Having multiple Gaussian processes is mostly useful for performance reasons. It is also relatively straightforward to build more complex model structures with several Gaussian processes that separate long-distance and short-distance contact.

The other point brought up by **Hammarström** is that we do not use AIC. This is a minor but important detail: AIC as well as Bayes factor have been shown to be suboptimal methods for comparison and evaluation of Bayesian models (and likely non-Bayesian as well). We provide full model comparisons using ELPD under cross-validation. According to Piironen and Vehtari (2017a), this is currently the standard for Bayesian statistics.

3.9 Chundra Cathcart

We thank **Cathcart** for his detailed but also accessible explanation of several models of feature evolution, as well as a discussion of some of their key differences. **Cathcart** suggests that there are perhaps alternative models of feature evolution that might be better suited for linguistic work. We do not have much to say here other than we do not doubt this. At the same time, we are unaware of systematic comparisons that could help us decide between the alternatives. Different researchers may not even agree with each other on how such comparisons should be carried out: through cross-validation on real data, simulations with synthetic data, or on theoretical principles? We hope that our study and this discussion can help to draw attention to this issue in quantitative typology, and that more work will be done to compare different modeling approaches so that we can reach a better understanding our modeling choices and their implications in the future.

3.10 Christophe Coupé

3.10.1 Model assumptions and model validation

Coupé discusses two important points in his commentary regarding model sanity checks, namely if the model assumptions are met and if the model does what we want it to do. The first one relates to something we did not mention explicitly, but which is always a crucial aspect of any quantitative analysis, namely that model likelihood choice must be justified by the data, and that models can give us incorrect answers if

their assumptions are not met. There are many more ways in which model assumptions could be violated, leading to incorrect results. We fully agree with **Coupé**, who states: "It is not common at all to see authors report that they have verified that the assumptions of their model(s) were met. This should become a standard requirement, as one among different sanity checks."

Coupé also suggests that researchers should present model diagnostics. We agree, with the caveat that there is no consensus yet as to what those diagnostics should be. Q-Q plots, as shown in **Coupé**'s commentary, appear to be rarely used in Bayesian statistics, with posterior predictive checks being seen as more insightful. Following mostly Vehtari et al. (2017), we have argued that cross-validation results are essential in evaluating model performance, and this is what we have used in this study as well as in other previous work. The practice of using cross-validation, however, seems to remain unpopular in linguistics, and not all models are easy to cross-validate (e.g. BayesTraits).

3.10.2 Better alternatives

Another point made by **Coupé** is that there might be better alternatives to phylogenetic regression and to the exponential quadratic kernel we used for the Gaussian process. This comment directly ties to our previous discussion on the complications of modeling space in Section 2.6, but also to **Cathcart**'s commentary on different models of feature evolution. Regarding kernel choices, we actually believe that Matérn family of kernels are a better choice.[13] We have also explored non-stationary Gaussian processes to model language expansion, with mixed results (Guzmán Naranjo et al. 2025; Urban and Guzmán Naranjo 2025).

As for phylogenetic effects, we strongly agree that they make simplified assumptions about feature evolution, which might not match very well how languages evolve. We are currently working on testing some alternatives. At the same time, we remain slightly skeptical of the following, specific comment by **Coupé**: "[…] the phylogenetic component in B&GN's models assumes that the rate of linguistic evolution is always the same along the different branches of the language tree. It is possible to relax this constraint/assumption (Davies et al. 2019), and therefore build a model that better matches the history of languages." We see two issues with this statement, one conceptual and one practical. The conceptual issue is that we do not know, for the most part, what a correct model, matching the history of language, would or should look like. We would be pleased to include more complex phylogenetic structures in our models, but we need to be cautious about those models not

13 The technical reason is that they produce less smooth areas with sharper transitions, which seem to better capture the geography of language.

only being more complex but potentially also introducing additional, wrong assumptions. In our models, we could implement different evolution rates by using independent phylogenetic terms for each family. This is straightforward from a modeling perspective, but the question is whether it makes sense from a data perspective, depending on how much we know about each of the language families included in the sample.

From a practical modeling perspective, we already find that phylogenetic regression can be too sensitive and too flexible in some cases. Prior choice has a considerable impact on the estimates (this actually relates to **Coupé**'s comment on the number of parameters vs. the number of data points), and wide priors lead to very poor performance under cross-validation. We also already mentioned that phylogenetic regression can fail, namely if the language sample contains many isolates or a high number of very small families with two to three languages. We are, of course, happy to be proven wrong on both points.

3.11 Matthew Dryer

Dryer's commentary focuses on our robustness test of his 2018 paper. Besides several important comments on the modeling of spatial relations between languages that we addressed in Section 2.6, **Dryer** emphasizes that when two methods produce (slightly) different results for the same dataset, it is crucial to try to understand how these differences came about. In particular, **Dryer** discusses two points of diverging results between his and our study. For the differences we found in the distribution of N-A-Num-Dem, **Dryer** agrees with our explanation of why his method potentially underestimates the cross-linguistic probability of this word order. For the word order Dem-Num-N-A, **Dryer** questions our explanation and offers an alternative one. We thank him for his clearly articulated correction and fully agree with the explanation. Finally, we want to emphasize **Dryer**'s general comment that "linguists should be wary of all typological studies using statistics until the results have been replicated". Of course, this also includes our own previous work.

3.12 Ilja Seržant

The commentary by **Seržant** mostly focuses on our robustness analysis of his 2021 paper. He raises a series of interesting theoretical issues as well as some criticism of our study. We reply to these points in turn. The issues brought up by **Seržant** make a great example of how the relation between the statistical analysis and the theoretical interpretation is not a trivial one and probably deserves more space in papers than it

often gets. In line with **Tallman**'s commentary, we take away from the discussion that we should all strive to make our theoretical assumptions, modeling decisions, our interpretations and conclusions as well the relations between these different moving parts as explicit as possible.

3.12.1 Explaining concrete decay factors

One of the main points mentioned by **Seržant** is that the model in our study does not explain the specific decay factors of Slavic languages. We agree to some extent. We do not have a complete, linguistic explanation of why the decay factors for Slavic are exactly what they are. Rather, our argument is that based on the data as presented by Seržant (2021) for the quantitative analysis, we cannot conclude that the specific values of Slavic are a result of contact with Turkic and Uralic, and that we do not find strong support for an East-West cline when adding phylogenetic and spatial components to our model. Seržant (2021) and **Seržant**'s commentary offer explanations for the decay factors of Slavic languages based on additional knowledge and evidence about the particular contact situation of Slavic languages in the past. This is very insightful from a theoretical perspective, but our methodological point is that we do not find evidence for the cline and the need for contact to account for the position of Slavic to begin with when analyzing the data.

3.12.2 Distinguishing between pressures and signals

Seržant makes a useful distinction between areal and phylogenetic pressures as mechanisms in the real world shaping language structures, and signals as found in statistical models. He points out that it is not always clear whether the relevant signals identified through modeling can necessarily be taken as evidence for the respective pressures to be relevant in the real world. Related to that, **Seržant** asks: "should a strong genealogical signal always be taken literally to mean that inheritance alone is sufficient, i.e., that there would be something specific to a family?" This is an excellent question and, unfortunately, a very difficult one to resolve. In models like the one we use, when there is complete overlap in the variance explained by two predictors, it is fundamentally impossible to decide categorically whether both predictors play a role in the real world or not. In this case, our result shows that a model which only includes the phylogenetic structure performs equally well as the model with a phylogenetic structure and areal structure. Does this mean we can discard any type of areal pattern? Not really, but it does mean that our certainty about the areal dynamics needs to be set much lower than if we found that both predictors together perform better than either predictor on its own.

3.12.3 Genealogical pressure and innovations

Related to the last point is the question of how we should interpret a phylogenetic bias in regard to decay factors. **Seržant** argues that "genealogical pressure is logically incompatible with innovations." Addressing this question properly would require a paper of its own, but we think that **Seržant**'s claim is potentially too strong. We can imagine that genealogy could bias a family branch towards innovation if a common change at some point in the past caused the system to be in an unstable or communicatively less-preferred state. Then, the languages belonging to that branch would have a greater than chance probability to undergo a change, leading to an innovation compared to the proto-language. The changes that individual languages would undergo could be different, but they could also be similar if there are universal pressures towards certain structures. In our opinion, it would therefore be possible that, at some point in the past, Romance and Germanic languages innovated some unstable configuration that resulted in more decay in their paradigms than what we observe in Slavic.

3.13 Olena Shcherbakova, Volker Gast, Simon J.Greenhill, Damián Blasi, Russell D. Gray and Hedvig Skirgård

The commentary by **Shcherbakova et al.** focuses on our robustness analysis of their 2023 study. Besides their points on terminology (cf. Section 2.5) and our choices for Euclidean distances (cf. Section 2.6), the authors raise four other issues that we will address in turn.

3.13.1 Inter-correlated evolution

One of the technical points mentioned by **Shcherbakova et al.** is that their model captures correlated evolution of two features, while ours "model[s] correlation between the variables that ignores the interdependence between the changes in these variables throughout time" and "[i]s a synchronic correlational analysis". We do not think that this opposition between the two approaches is entirely accurate because our phylogenetic term also captures correlated feature evolution, and we do not completely agree with the characterization of our approach to be entirely synchronic. Our model explicitly samples two phylogenetic terms (which capture diachronic evolution based on a specific assumption of diachronic development), but samples them as correlated, with an explicit correlation parameter in the model. The correlation is not an a posteriori metric, and it is not dependent on how well the

model can predict the dependent variables. We admit that we could have made this point more explicit in our study.

One can, of course, argue that phylogenetic terms are not a good representation of linguistic evolution (see **Cathcart**'s commentary for some discussion), or that BayesTraits is fundamentally better. However, as far as we are aware, these are currently unknowns. We have no solid data about which types of models work best with what types of situations. The fact that BayesTraits or our model yield some result does not mean that the result is correct. We did try to evaluate our model against synthetic data for which we knew the real parameter values, and we could recover the parameter values very well. In their commentary, **Shcherbakova et al.** argue that our simulated data is not a good representation of real linguistic data. They may be correct, but this is yet another open question that needs to be tested empirically. One of the points of our paper was precisely to show how much uncertainty there currently is regarding methods in quantitative typology including computational models, and how this calls for more caution with any single result.

3.13.2 Contact versus no contact

Shcherbakova et al. note that one of the differences between their and our approach is that we include spatial dependencies as a proxy for contact between the languages in the model. The authors argue that this "further widens the divide between the two studies' theoretical assumptions" and makes it less of a replication study. As we mentioned in Section 2.5, we gladly follow **Shcherbakova et al.**'s suggestion to classify our study as a robustness analysis. As such, we do not understand why it is an issue that we include spatial effects. We think that it should be desirable in this case to include some form of control for contact in our models, whether evolutionary or regression based, as we can assume that contact between languages will have played a role in at least some of the morphosyntactic changes analyzed. If a model does not include any control for potential contact effects, then we cannot possibly know whether the results really represent language-internal evolutionary dynamics alone, or whether contact dynamics could have also impacted the changes. In other words, we need to control for confounding factors.[14]

3.13.3 Multicollinearity

Shcherbakova et al. are concerned about multicollinearity between our predictors, as we predict the outcome of the 13 feature groups simultaneously instead of making

14 Note that it is also possible to use a phylogenetic framework similar to BayesTraits with a contact component (cf. Hartmann and Jäger 2024).

pair-wise comparisons like Shcherbakova et al. (2023). Multicollinearity is an issue which appears in a (generalized [additive]) linear model of the form $y \sim F(\alpha + \beta_1 x_1 + \beta_2 x_2, \dots)$ where F is some likelihood function, and x_2 and x_2 are correlated predictors. This leads to a situation where estimating β_1 and β_2 becomes difficult because there is a wide range of values which work equally well, and thus we cannot reach a good estimate for the real values of the coefficients. There are two points to our reply regarding this concern. Our first point is that under a Bayesian framework, multicollinearity – if it does exist – can be mitigated considerably and even overcome by careful prior choice (Al-Essa et al. 2024). The reason is that if we provide the model with a good guess about likely values of β_1 and β_2, then the model has an easier time finding reasonable estimates for both. Of course, there could still be issues of high uncertainty, but it is a much less pronounced problem than with maximum likelihood estimation. Other alternatives like Dirichlet processes or Horseshoe priors can also be used in the presence of multicollinearity (cf. Piironen and Vehtari 2017b).

Second, and more importantly regarding our study, there is no such risk in our model setup because we do not use the individual variables as predictors but as dependent variables in the model. Hence, there is no question of collinearity. The correlation comes from the fact that we sample the different phylogenetic components from a matrix normal distribution, which means that we sample them as correlated (there is an explicit correlation coefficient in the model).

3.13.4 Interpretation of the results

In their commentary, **Shcherbakova et al.** also discuss differences between our and their interpretation of how results compare. The authors state that they do not think that the results of Shcherbakova et al. (2023) and those of our study are as different as we make them out to be. They note that they find a null result and we find a positive correlation between nominal and verbal complexity, which are both in agreement that there is no trade-off (which would imply a negative correlation). We do not think that we actually disagree that much, as we wrote the following: “While we could not fully replicate Shcherbakova et al.’s results, we did not find entirely opposite trends and patterns either” (Becker and Guzmán Naranjo 2025). At the same time, the models produce different effect sizes and effect directions (no effect vs. positive correlation) in some cases, which is what we highlighted as interesting. We think that it is worth trying to understand where these differences come from, but we do not have a good answer in this case. Perhaps **Tallman** is correct in pointing out that a stronger theory, i.e. a less exploratory conceptual setup could help in designing more specific and less complex statistical models whose results and comparisons could be easier to interpret.

References

Al-Essa, Laila A., Endris Assen Ebrahim & Yusuf Ali Mergiaw. 2024. Bayesian regression modeling and inference of energy efficiency data: The effect of collinearity and sensitivity analysis. *Frontiers in Energy Research* 12. 1–20.

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell & Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics (Version 1). Research Data Alliance. https://doi.org/10.15497/RDA00040.

Becker, Laura, Matías Guzmán Naranjo & Samira Ochs. 2023. Socio-linguistic effects on conditional constructions: A quantitative typological study. In Silvia Ballarè & Guglielmo Inglese (eds.), *Sociolinguistic and typological perspectives on language variation*, 121–154. Berlin: De Gruyter.

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

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. 2018a. The Austin principles of data citation in linguistics (Version 1.0). https://site.uit.no/linguisticsdatacitation/austinprinciples (accessed 15 June 2025).

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. 2018b. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18.

Bisang, Walter & Andrej Malchukov (eds.). 2020a. *Grammaticalization scenarios from Africa, the Americas, and the Pacific*. Berlin: De Gruyter.

Bisang, Walter & Andrej Malchukov (eds.). 2020b. *Grammaticalization scenarios from Europe and Asia*. Berlin: De Gruyter.

Bolstad, William & James Curran. 2017. *Introduction to Bayesian statistics*. Hoboken: Wiley.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097). 957–960.

Coretta, Stefano, Joseph V. Casillas, Simon Roessig, Michael Franke, Byron Ahn, Ali H. Al-Hoorie, Jalal Al-Tamimi, Najd E. Alotaibi, Mohammed K. AlShakhori, Ruth M. Altmiller, Pablo Arantes, Angeliki Athanasopoulou, Melissa M. Baese-Berk, George Bailey, Cheman Baira A. Sangma, Eleonora J. Beier, Gabriela M. Benavides, Nicole Benker, Emelia P. BensonMeyer, Nina R. Benway, Grant M. Berry, Liwen Bing, Christina Bjorndahl, Mariška Bolyanatz, Aaron Braver, Violet A. Brown, Alicia M. Brown, Alejna Brugos, Erin M. Buchanan, Tanna Butlin, Andrés Buxó-Lugo, Coline Caillol, Francesco Cangemi, Christopher Carignan, Sita Carraturo, Tiphaine Caudrelier, Eleanor Chodroff, Michelle Cohn, Johanna Cronenberg, Olivier Crouzet, Erica L. Dagar, Charlotte Dawson, Carissa A. Diantoro, Marie Dokovova, Shiloh Drake, Fengting Du, Margaux Dubuis, Florent Duême, Matthew Durward, Ander Egurtzegi, Mahmoud M. Elsherif, Janina Esser, Emmanuel Ferragne, Fernanda Ferreira, Lauren K. Fink, Sara Finley, Kurtis Foster, Paul Foulkes, Rosa Franzke, Gabriel Frazer-McKee, Robert Fromont, Christina García, Jason Geller, Camille L. Grasso, Pia Greca, Martine Grice, Magdalena S. Grose-Hodge, Amelia J. Gully, Caitlin Halfacre, Ivy Hauser, Jen Hay, Robert Haywood, Sam Hellmuth, Allison I. Hilger, Nicole Holliday, Damar Hoogland, Yaqian Huang, Vincent Hughes, Ane Icardo Isasa, Zlatomira G. Ilchovska, Hae-Sung Jeon, Jacq Jones, Mágat N. Junges, Stephanie Kaefer, Constantijn Kaland, Matthew C. Kelley, Niamh E. Kelly,

Thomas Kettig, Ghada Khattab, Ruud Koolen, Emiel Krahmer, Dorota Krajewska, Andreas Krug, Abhilasha A. Kumar, Anna Lander, Tomas O. Lentz, Wanyin Li, Yanyu Li, Maria Lialiou, Ronaldo M. Lima, Jr., Justin J. H. Lo, Julio Cesar Lopez Otero, Bradley Mackay, Bethany MacLeod, Mel Mallard, Carol-Ann Mary McConnellogue, George Moroz, Mridhula Murali, Ladislas Nalborczyk, Filip Nenadić, Jessica Nieder, Dušan Nikolić, Francisco G. S. Nogueira, Heather M. Offerman, Elisa Passoni, Maud Pélissier, Scott J. Perry, Alexandra M. Pfiffner, Michael Proctor, Ryan Rhodes, Nicole Rodríguez, Elizabeth Roepke, Jan P. Röer, Lucia Sbacco, Rebecca Scarborough, Felix Schaeffler, Erik Schleef, Dominic Schmitz, Alexander Shiryaev, Márton Sóskuthy, Malin Spaniol, Joseph A. Stanley, Alyssa Strickler, Alessandro Tavano, Fabian Tomaschek, Benjamin V. Tucker, Rory Turnbull, Kingsley O. Ugwuanyi, Iñigo Urrestarazu-Porta, Ruben van de Vijver, Kristin J. Van Engen, Emiel van Miltenburg, Bruce Xiao Wang, Natasha Warner, Simon Wehrle, Hans Westerbeek, Seth Wiener, Stephen Winters, Sidney G.-J. Wong, Anna Wood, Jane Wottawa, Chenzi Xu, Germán Zárate-Sández, Georgia Zellou, Cong Zhang, Jian Zhu & Timo B. Roettger. 2023. Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science* 6(3). 1–29.

Cysouw, Michael. 2007. A social layer for typological databases. In Andrea Sansò (ed.), *Language resources and linguistic theory*, 59–66. Milan: Franco Angeli.

Davies, T. Jonathan, James Regetz, Elizabeth M. Wolkovich & Brian J. McGill. 2019. Phylogenetically weighted regression: A method for modelling non-stationarity on evolutionary trees. *Global Ecology and Biogeography* 28(2). 275–285.

Desagulier, Guillaume. 2017. *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Berlin: Springer.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68(1). 81–138.

Dryer, Matthew S. 2013a. Order of object and verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: https://wals.info/chapter/83.

Dryer, Matthew S. 2013b. Order of subject, object and verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: https://wals.info/chapter/81.

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

Egbert, Jesse, Douglas Biber, Bethany Gray & Tove Larsson. 2025. Achieving stability in corpus-based analysis of word types. *International Journal of Corpus Linguistics*. aop. https://doi.org/10.1075/ijcl.24109.egb.

Flanagan, Joseph. 2025. Reproducibility, replicability, robustness, and generalizability in corpus linguistics. *International Journal of Corpus Linguistics*. aop. https://doi.org/10.1075/ijcl.241.

Fried, Eiko. 2020. Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry* 31(4). 271–288.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics*, 91–100. Uppsala: ACL.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2013. *Bayesian data analysis*, 3rd edn. Boca Raton: CRC Press.

Gelman, Andrew & Eric Loken. 2014. The statistical crisis in science. *American Scientist* 102(6). 460–466.

Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2019. Rediscovering Greenberg's word order universals in UD. In *Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 124–131. Stroudsburg, PA: The Association for Computational Linguistics.

Gerdes, Kim, Sylvain Kahane & Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa* 6(1). 1–17.
Greenberg, Joseph. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg (ed.), *Universals of language*, 73–113. Cambridge, MA: MIT Press.
Gries, Stefan. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge.
Gries, Stefan. 2013. *Statistics for linguistics with R: A practical introduction*. Berlin: De Gruyter.
Gries, Stefan. 2025. Closing remarks and outlook. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.24120.gri (Epub ahead of print).
Grieve, Jack. 2021. Observation, experimentation, and replication in linguistics. *Linguistics* 59(5). 1343–1356.
Guzmán Naranjo, Matías & Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, 91–104. Linköping: Linköping University Electronic Press.
Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.
Guzmán Naranjo, Matías, Laura Becker, Miriam L. Schiele & I-Ying Lin. 2025. Why modelling space is hard: No evidence for a serial founder effect in Polynesian phoneme inventories. *Linguistics*. https://doi.org/10.1515/ling-2024-0016 (Epub ahead of print).
Guzmán Naranjo, Matías & Gerhard Jäger. 2023. Euclide, the crow, the wolf and the pedestrian: Distance metrics for linguistic typology. *Open Research Europe* 3. 104.
Guzmán Naranjo, Matías & Miri Mertner. 2023. Estimating areal effects in typology: A case study of African phoneme inventories. *Linguistic Typology* 27(2). 455–480.
Guzmán Naranjo, Matías, Miri Mertner & Matthias Urban. 2024. Spatial effects with missing data. *Open Linguistics* 10(1). 20240032.
Hartmann, Frederik. 2022. Methodological problems in quantitative research on environmental effects in phonology. *Journal of Language Evolution* 7(1). 95–119.
Hartmann, Frederik & Gerhard Jäger. 2024. Gaussian process models for geographic controls in phylogenetic trees. *Open Research Europe* 3. 57.
Hartmann, Frederik, Seán Roberts, Paul Valdes & Rebecca Grollemund. 2024. Investigating environmental effects on phonology using diachronic models. *Evolutionary Human Sciences* 6. e8.
Hawkins, John. 1983. *Word order universals and their explanation*. New York: Academic 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.
Laitinen, Mikko & Paula Rautionaho. 2025. Reuse of social media data in corpus linguistics. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.24136.lai (Epub ahead of print).
Lamport, Leslie. 2009. The PlusCal algorithm language. In *International colloquium on theoretical aspects of computing*, 36–60. Berlin: Springer.
Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: Benjamins.
Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.
Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova & Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825–883.

Malchukov, Andrej & Bernard Comrie (eds.). 2015a. *Valency classes in the world's languages: Volume 2. Case studies from Austronesia, the Pacific, the Americas, and theoretical outlook*. Berlin: De Gruyter.

Malchukov, Andrej & Bernard Comrie (eds.). 2015b. *Valency classes in the world's languages. Volume 1. Introducing the framework, and case studies from Africa and Eurasia*. Berlin: De Gruyter.

McElreath, Richard. 2020. *Statistical rethinking*, 2nd edn. Boca Raton, FL: CRC Press.

Merkel, Dirk. 2014. Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal* 2014(239). 2.

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

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

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

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

Östling, Robert. 2015. Word order typology through multilingual word alignment. In Chengqing Zong & Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 205–211. Beijing, China: Association for Computational Linguistics.

Piironen, Juho & Aki Vehtari. 2017a. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* 27. 711–735.

Piironen, Juho & Aki Vehtari. 2017b. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 905–913. Fort Lauderdale: PMLR.

Riutort-Mayol, Gabriel, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin & Aki Vehtari. 2023. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Statistics and Computing* 33(1). 17.

Roberts, Seán, James Winters & Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. *PLoS One* 10(7). e0132145.

Roettger, Timo. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology* 10(1). 1–27.

Roettger, Timo, Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.

Roettger, Timo. 2021. Preregistration in experimental linguistics: Applications, challenges, and limitations. *Linguistics* 59(5). 1227–1249.

Schweinberger, Martin & Michael Haugh. 2025a. Reproducibility and transparency in interpretive corpus pragmatics. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.23033.sch (Epub ahead of print).

Schweinberger, Martin & Michael Haugh. 2025b. Reproducibility, replicability, and robustness in corpus linguistics: An introduction. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.25081.sch (Epub ahead of print).

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 E. Blasi, Hedvig Skirgård, Russell D. Gray & Simon J. 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.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill,

Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.

Smith, Kenny. 2024. Simplifications made early in learning can reshape language complexity: An experimental test of the linguistic niche hypothesis. *Proceedings of the Annual Meeting of the Cognitive Science Society* 46. 1346–1352.

Sönning, Lukas & Valentin Werner. 2021. The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59(5). 1179–1206.

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

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

Talamo, Luigi & Annemarie Verkerk. 2022. A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics* 34(2). 171–226.

Urban, Matthias & Matías Guzmán Naranjo. 2025. Gradient in grammatical structure of indigenous languages reflects pathway of human expansion in the americas. *Nature Scientific Reports* 15(14365). 1–15.

Van Tuyl, Rory & Asya Pereltsvaig. 2012. Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". *Science* 335(6069). 657.

Vasishth, Shravan & Andrew Gelman. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* 59(5). 1311–1342.

Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.

Wichmann, Søren & Taraka Rama. 2021. Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1824). 20200202.

Winter, Bodo. 2020. *Statistics for linguists: An introduction using R*. New York: Routledge.