Review of "Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research"

This submission is an overview of the important steps towards open science and transparent research in linguistics. In offering detailed motivations of why these steps towards open science are important, it is a very timely and suitable contribution to the *Journal of Linguistics*. In general, the structure and the argumentation of this submission are very clear and easy to understand. Besides a number of detailed comments that I list below, I had two general issue reading this manuscript, which should be easy to fix:

- The argumentation and description of methods is generally very abstract. If I understand the title "opening open science to all" correctly, the authors should include more concrete examples and instructions so that readers who are (still) unfamiliar with the practices described can implement new methods. I added concrete suggestions to do so below; this is necessary for the manuscript to fulfill its promise of opening open science and to become an important resource in research and teaching.
- This manuscript is not aimed at research from any particular linguistic area. Still, it leaves the reader wondering in many places which areas and methods certain arguments and steps apply to, as well as how certain steps could be applied to linguistic research outside of experimental approaches. I added concrete suggestions below; I suggest that the authors either mention more explicitly which areas / methods certain steps apply to and how, or that they restrict the manuscript to "transparency practices in experimental / empirical / ... linguistic research".

Detailed comments:

Abstract

- ERC: should be spelled out here for clarity, as it is only explained later in the introduction.

Introduction

- "Researchers have pointed to questionable research practices (QRPs), such as p-hacking and HARKing"
  --> HARKing (maybe also p-hacking) needs to be explained briefly here, especially given the goal of the paper.
- "It necessitates that researchers implement new techniques with limited pedagogical resources and embrace alternative methods of disseminating their research, all of which constitutes a steep learning curve."
  --> Does it always, though? For instance, in the field of language typology, there are a number of qualitative studies that are based on a sample and annotations using reference grammars. Those studies do not include any code, making them transparent simply means providing a spreadsheet with the languages and the respective annotations and sources that the authors should have in some form anyway. Can we really speak of "innnovative methodologies"?All this goes to say that, yes, some formats (e.g. using OSF) may involve a learning curve, but this is not necessarily so. It may also be about vulnerability, i.e. making all details of the study and analysis public and thus subject to potential criticism, wheareas before, those could be left partially "hidden" and "proctected". I suggest that the authors sightly rephrase this paragraph.
- "To this end, we identify three areas, stance, workflow, and dissemination, in which linguists can engage in open science."
  --> rephrase: "... we identify the following three areas of stance, workflow and dissemination, ...

- Figure 1 --> I am not sure open data is clearly only related to stance; I would also view providing the datasets in a useful form as part of the workflow. Could this point be put in between the two areas?

Positionality statements

- Could you give a concrete example of a positionality statement early on? A concrete example is valuable to make this more understandable for researchers not familiar with this practice.
- "Moreover, Bucholtz et al. (2023) note that considering a researcher's positionality may be especially important in linguistics, "[…] which relies on racially minoritized communities as sources of data yet lack adequate (if any) representation of those communities among faculty researchers" (p. 2)."
  --> Does this quote really apply to linguistics in general? I can see how it is very relevant for certain areas of linguistics, but there is much linguistic research on major languages, where this point does not seem to apply. Please rephrase / explain.
- "Many researchers support and advocate for the inclusion of positionality statements in their research publications (e.g., Bucholtz et al., 2023; Jafar, 2018; Steltenpohl et al., 2022)."
  --> Please specify the research areas here, also those mentioned as arguing against positionality statements.
- "Bochynska et al. (2023) surveyed open and transparent practices in linguistics and found that only 10% of the articles sampled included statements of conflict of interest, and, among those 10%, none declared any conflicts (See also Cristea & Ioannidis, 2018; Hardwicke et al., 2022, 2020)."
  --> Out of curiosity, can you give an example of a statement of conflict of interest from linguistics? Again, concrete examples would help readers less aware of / familiar with these issues understand what those statements can look like.
- "While positionality statements, due to their reflexive nature, may encompass larger pieces of writing, they can also take the form of short paragraphs that illustrate a few personal characteristics deemed relevant for the particular research endeavor."
  --> This whole paragraph should be moved to the beginning of the section.
- "Gabriela is a white immigrant cis-gender woman from Romania whose research focuses on how non-native speakers are ideologically framed as linguistically deficient in comparison to native speakers who are characterized by their linguistic authority and expertise."
  --> OK, this is what I wanted to see early on in this section. Please add some context: What was the title / topic of the publication that this statement belongs to? Can you give a concrete reference? Or is this a made-up example? Please clarify.
- "For examples of positionality statements in linguistic research, the interested reader is directed to Bochynska et al. (2023) and Weissler et al. (2023)."
  --> Great pointer to ressources, but if the purpose of this paper really is to make open science practices more accessible to linguists who are not (yet) familiar with them, a concrete example in the beginning of this section is important. It would also help to make this proposal of including a positionality statement more concrete: In which areas of linguistics is it helpful? In which areas is it necessary? In which areas may it be less relevant? For which methods may this be particularly helpful / important? While I sympathize with the arguments made in this section in general, I think they are too abstract in order to be immediately helpful for a concrete implementation. Please make the argumentation a bit more concrete and geared towards certain / different areas of linguistic research so that the reader has more practical advice on how to include such statements in their future research.

Open data

- “In academic research, statements such as “data available upon request” are commonplace.”
  --> This may be so, but the statement would clearly look better with references or some data to back it up. I do not mean that the authors should single out studies as negative examples, but maybe it would be possible to mention journals and years of publication where this happens. This makes it somewhat more tangible and related to linguistic research. Please rephrase.
- “Recent efforts have attempted to encourage researchers to make linguistic data open and accessible via servers (e.g., the IRIS database).”
  --> Given the goal of this paper, please add more explanation about the IRIS database here.
- “open science bagdes” --> Can you explain briefly what this is?
- “Though researchers may understandably hesitate to share their data, we believe understanding the benefits of open data can help alleviate any concerns.”
  --> This point needs to be elaborated more in that there may be very different motivations for preferring not to share data. Corpus data may contain sensitive personal information (e.g. video, audio) and speakers may not consent to making it public. Authors of the study may not have the rights to publish the data that their study is based. In this case, it may be possible to make processed data publicly available, but not the raw data (i.e. data can mean different things, and should be distinguished in more detail here). In other cases, it may simply be the fear of making all data and annotation choices criticiseable by making everything publicly available. Those are very different perspectives that call for very different solutions / approaches. This needs to be discussed and differentiated in more detail.
- “It affords third parties the opportunity to scrutinize original findings, which promotes reproducibility and reduces errors, such as those related to statistical analyses and reporting of outcomes [TIMO].” --> What is [TIMO]?
- “Revisiting old data sets using innovative techniques can support or contradict past narrative conclusions (e.g., Casillas, 2021).”
  --> It would be useful to elaborate this in a few sentences and present the Casillas (2021) study in terms of revisiting old data sets and drawing (new?) conclusions.
- “Open data are particularly important for the field of linguistics, for all of the aforementioned reasons, and also because linguistics is Western, Educated, Industrialized, Rich, and Democratic (WEIRD, Bochynska et al., 2023).”
  --> I do not disagree, in principle, but there may be one caveat that needs to be mentioned: Bochynska et al. (2023) evaluated publications written in English. This seems slightly self-selecting to me, as it probably excludes relevant linguistic research traditions whose main language of publication is not English (e.g. Chinese linguistics). How sure can we then be that linguistics as a whole is WEIRD? I suggest the authors tone down this statement a bit, as it seems too strong in its current form.
- “Making linguistic data accessible to all researchers promotes participation *in* and *with* underrepresented communities. Furthermore, it can increase the study of diverse and underreported languages, which fosters a more inclusive and comprehensive understanding of the global linguistic landscape.”
  --> Please explain why and how accessible linguistic data promotes the researchers’ participation with the (potentially underrepresented) speaker communities? While I am sympathetic to this argument, it needs to be spelled out more and made more concrete in order to be meaningful.
- “Having stated all the above, it is necessary to recognize that the field of linguistics faces unique challenges with regard to open data.”
  --> Are these necessarily unique challenges? I do not think it is important to point this out (and I am not sure if linguistics really faces unique challenges that different from other related disciplines). What would be more important is to go into more detail about which areas of linguistics face which types of challenges, since there are so many different types of

data that we as linguists work with (experimental data, corpus data, data from grammars, data from linguistic databases, written, spoken, signed data, historical data, and each oft these types could be distinguished further). This should be discussed in much more detail and made more concrete so that the readers can see their approaches / types of data represented in this discussion.

- "The privacy and consent of participants must be safeguarded..."
  --> OK, this is the kind of information that I was waiting for earlier; consider moving this paragraph up. Also, not all linguistic studies face ethical issues: What about studies that work with databases / corpora that have been compiled by third parties, which are publicly accessible and which the authors simply use as is. They are far removed from any speakers in this case and had no influence on any decision of data collection and processing that went into compiling the database. Or what about historical linguistics? The role that participants or informants play in different types of linguistic studies deserves a more fine-grained discussion.
- "Moreover, the advent of generative artificial intelligence technologies, such as Large Language Models, may pose unknown challenges in the near future that necessitate additional steps to secure the protection of sensitive data against misuse."
  --> Please explain in more detail: Is the idea that LLMs use data compiled by linguists for linguistic research? Or does the potential of misuse lie in the linguistic research with LLMs itself? If the authors wish to mention this, they should discuss in more detail what exactly the risks are. Otherwise, this sentence could be omitted.
- "When primary data, such as audio or video files, cannot be shared, derived data in the form of tabular files can take its place..."
  --> This paragraph is very important and constructive. Different types of data (primary data vs. processed data and annotations) should be introduced at the beginning of this section, though, and the entire section should be more precise about what types of data particular arguments refer to.
- Prolific --> add a link, e.g. in a footnote.
- "In more uncommon cases in which institutional policies do not permit the sharing of derived data sets, synthetic data containing the same statistical properties can be generated and shared freely (See Quintana, 2020)."
  --> Please elaborate how this would work.
- "A substantial hurdle that cannot be overlooked revolves around the fact that researchers must learn to use new technologies to participate in open, transparent research. Making data open and accessible is not as simple as merely uploading a data file."
  --> I disagree in that it depends on the linguistic subdiscipline. In large-scale typological studies that are qualitative in nature and use no statistics but are based on a language sample, making research transparent simply consists of adding a supplementary spread sheet file with the names of the languages, all relevant annotations / examples and their sources. Impressionistically, even this is still not done in all studies, so improving on transparency in this case only involves adding a supplementary file to the publication. Also, if the researcher does not provide such supplementary files on their own, many journals offer to host supplementary materials on the journal website together with the publication. This may not be the best solution, but it is a solution where the researcher really does not need to know or consider platforms to host data. Please clarify this.
- "Free repositories designed for the purpose of sharing research materials, such as the Open Science Framework (Foster & Deardorff, 2017), github, etc., are preferable and can be accessed simply by sharing a link."
  --> Here, it would actually be helpful to spell this out a bit more and mention e.g. zenodo as another option. For readers less familiar with these options, it would be valuable to add a table with different platforms / systems and provide information about long-term support,

version control, possibility of DOI, anonymous links (e.g. what OSF offers and what is very useful for sharing data and code during the revision stage of a study), etc.

- "While linguistics does face legitimate, field-specific challenges, ultimately the benefits of open data outnumber these challenges, and researchers should take the stance to share what is ethically reasonable."
  --> My impression when reading this is that the only field-specific challenge mentioned was the point on sensitive speaker data. I generally agree with the statement, but in order to make this statement at the end of the section, the section would need to be much more explicit about the different challenges taking into account the different areas of linguistics.
- Another point to consider is the following, although it may not immediately relate to data and more to the inclusion of lesser studied languages and integration of speaker communities in linguistic research: There are linguistic journals that allow / require an additional abstract of the papers in the target language or another local language besides the usual abstract in English. From personal experience, I know that Linguistics Vanguard offers this as an option (although I am not sure if you can find this information other than submitting a paper). Another journal that makes use of this practice is the Journal of African Languages and Linguistics (just to give some examples from the most recent volume: https://www.degruyter.com/document/doi/10.1515/jall-2023-2008/html, https://www.degruyter.com/document/doi/10.1515/jall-2023-2011/html)

Reproducibility

- "As we have seen, reproducibility is now a crucial aspect of any scientific study."
  --> Where did we see that? Also, I agree that it should be, but reproducibility is not yet a crucial aspect of any scientific study, unfortunately. Please rephrase.
- "At worst, a lack of reproducibility can lead to irreproducible results and wasted resources."
  --> Is this really the worst consequence? Is it not worse (or equally bad) that we may end up with results that are taken as a given for decades in a given linguistic field, because the original study was not reproducible and never replicated but influential for some reason?
- "This can have serious implications for public health and policy decisions based on research findings."
  --> Is this statement about linguistics or research in general? Please clarify.
- "In linguistics there is increasing awareness of the importance of reproducibility, and many researchers are beginning to take steps to improve the reproducibility of their research."
  --> too many reproducibilities in one sentence, rephrase.
- "There are several steps that researchers can take to make their code and projects more reproducible."
  --> A large portion of linguistic work is qualitative and does not rely on any code. If the paper is meant to be relevant to all linguistic approaches in general, please be more inclusive. Alternatively, the topic of the paper could be specified more and refer to "quantitative linguistics"?
- regarding the sharing of code, dependencies, versions & reproducibility: It may be helpful to explain in more detail why it is important to not only share the code, i.e. that different operating systems can react to code differently, functions can differ in different versions of packages, etc. The authors may consider referring to Roberts et al. (2015), who show that functions from the lme4 R package produce different results when used on different operating systems (althouh fixed since then).
  Roberts, Seán, James Winters & Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. *PLOS ONE* 10(7). e0132145. https://doi.org/10.1371/journal.pone.0132145.
- The literate coding section is very much focused on R. This may be warranted given that R seems to be the main programming language used for statistical analyses in many (?) linguistic areas, but python is certainly not uncommon either. If the authors go into detail for

literate programing in R (e.g. citing the knitr package), they should also do this for python, and potentially mention Julia, which is also, although less commonly, used in linguistics. Here again, a table with programing languages, approaches / formats / packages as well as platforms with virtual environments would be very helpful.

- "This includes updating code and documentation as needed and testing projects on different operating systems to ensure that it can be run in different environments."
  --> This statement seems to assume that researchers should have access to different OSs, which may not always be the case. Also, it is hardly possible to test for all compatibilities and eventualities on the side of the authors. It is more reasonable to create and publish an, e.g., Docker environment for an R project, so that other researchers can reproduce the original analysis with the exact OS and package versions as in the original study.

Preregistration

- "Linguistic research is multifaceted and spans diverse areas such as corpus analysis, conversation/discourse analysis, experimental research, and more."
  --> This is a very reduced list of linguistic research areas, please add more areas and discuss in more detail for which areas preregistration makes sense. For instance, can it be used in theoretical linguistics? It is mentioned in the beginning that "we focus on who might want to consider preregistrations", but this does not become very clear. Please add at least a paragraph about what types of linguistic studies can benefit from preregistrations.
- "Researchers face vital decisions while engaging in research, with inherent flexibility involved in the process of designing and conducting experiments, as well as analyzing the results (Simmons, Nelson, & Simonsohn, 2011)."
  --> Does this mean you only consider experimental linguistics research? This needs to be framed differently, or, the scope of the paper should be reduced to experimental linguistics.
- Coretta et al. 2023: Please explain in a few sentences what the study did and what it showed; it is impossible to understand (and appreciate) that without knowing the study.
- This section needs more concrete pointers to specific platforms, journals, spaces for preregistration. Linguists outside of experimental research may want to do this, but are less likely to know how and where exactly they can preregister their study. The abstract motivations are spelled out clearly in this section, but it is not of much concrete help to those linguists who are convinced about the method but have no experience with it. It would be helpful to see an example of how this can be done outside of experimental studies, e.g. a corpus study, a typological study, etc. What about e.g. a theoretical morphological / phonological / syntactic analysis based on information from grammars? Is preregistration even possible or sensible? Please add more concrete details regarding different areas of linguistics so that this section becomes usefull to the readers who are not (yet) familiar with this tool.

Registered reports

- Please give examples of journals / platforms that allow for registered reports.

Pre-prints

- "First, select a pre-print server that aligns with the course of research."
  --> Please give examples of suitable pre-print servers for different linguistic areas. Again, a table would be very helpful.

Concluding remarks

- "We have provided descriptions and relevant examples of these practices to accompany the many guides already available for learning open science (e.g., Crüwell et al., 2018; Lewis, 2020). Crucially, the purpose of this article is to help foster open science in linguistics (FOSIL)."

--> The current version of this manuscript provided detailed motivations and descriptions, but not sufficient examples and concrete instructions for linguists who are less familiar with these practices. Please add more concrete details, examples, instructions as indicated above.
--> The FOSIL tutorials should be mentioned much earlier whenever relevant, given that they do include more concrete information.

References

- Some references contain incomplete author lists --> all authors should be cited.