October 8, 2019

Dear Editors,

We would like to thank you and the reviewers for your very thoughtful comments on our DMUC. We have revised it thoroughly in response. We left “Track changes” on in the Word document as we edited it, so you should be able to see a direct record of the changes in the chapter .docx that we uploaded. Here are the issues you highlighted in your editorial letter (our responses in blue):

> Reviewer A has two main considerations for you. The first is regarding framing of the paper. When we asked you to write a DMUC, we suggested a title for you, but A points out that the title may not be the best one for the (excellent) DMUC you have written. If you would like to take his comments to heart and reframe the paper with a new title, that would be fine with us. And his comments about acknowledging that the paper is about a specific flavor of semantics is also important for properly framing the paper. In other words, the paper you wrote is a good one, it just needs to be situated better.
>
> We have changed the title to the suggested one, “Managing web experiments for psycholinguistics: An example from experimental semantics/pragmatics” and have added prose mostly to the end of section 1 to indicate the generalizability of the framework to other types of studies.
>
> A's second consideration has to do with the flow of the paper. I'll let you read his suggestions yourself, but in general I agree that a change in the order of presentation would make the paper more accessible to nonspecialist (like me); and there is some jargon that needs to be explained. The audience will be a mix of specialists, non-specialists (in semantics) and students, so it's important to be able to reach all of them.
>
> We have tried throughout to provide more motivation for the particular workflow we use. We have also tried to make section 1 slightly less jargon-y, or rather, to elaborate slightly on the jargon.
>
> Reviewer B also likes this paper very much and has some specific comments for you to address. B tells me that the first set of comments ("Minor but essential

issues") are more crucial than those in the second set of comments ("other thoughts and recommendations").

We have implemented their suggestions (see point by point responses below).

In addition to the suggestions from the reviewers, I would also ask that you ensure that every tool/website you mention gets either a URL or a reference at first mention (eg R, git, github), and please use the full URL in all cases (eg submiterator). It's fine to put URLs in footnotes if the body of the paper gets too cluttered up with them. I also would like to ask that you try to find places to reference Part 1 chapters.

All tools and websites now have their URLs listed in footnotes, and we have included an overview table of all the tools we used (Table 1). We have also referred to chapters 1, 2, and 5 for the general spirit of open science, though without seeing the full chapters it's hard to know which references to include.

Below you'll find a prose description of all the changes as responses to your and the reviewers' comments. We hope that you will like the new version of the DMUC. Please let us know if you require additional information, and we look forward to hearing from you!

Sincerely,

Judith Degen and Judith Tonhauser

Responses to reviewers' comments

Reviewer A:

The reviewer self-identifies to be Bodo Winter and is available for follow-up questions.

This is a useful paper and it will definitely be a great contribution to the handbook. I'd judge the quality of the data management use cases by whether they are able to teach me something new, and that was definitely the case here. I run a lot of web experiments but wasn't aware of some of the tools mentioned in this manuscript (supersubmiterator, Unique Turker). The paper is also well written.

My main issues with the paper are two-fold:

1) First, the framing of the overall paper.
2) Second, the structure.

Both of these things can be addressed rather easily by re-writing the introduction a little bit and perhaps by reshuffling some of the sections. Not too much needs to be changed, just a bit more signposting and changing a few opening sentences to make the whole thing appear in a different light.

1) Reframing
The paper is called "Managing data for experimental semantics and pragmatics" but, really, the paper is about how to run web experiments in a reproducible fashion. I think a change in title and in focus is called for. I'd perhaps call it "Managing web experiments for psycholinguistics: An example from experimental pragmatics" or so. I find the current title really misleading and I think that the paper will get more recognition and wider citations if it is pitched accordingly. It certainly is the case that the methods here applicable to many domains of linguistics (including experimental syntax, experimental sociolinguistics etc.). So why constrain the focus so much and limit the readership?

We have now changed the title according to Bodo's suggestion.

On top of that, it has to be acknowledged that what the authors mean by "experimental semantics" is a specific type of semantics, one closely aligned with the formal tradition. In fact, I have been involved in a handbook chapter on "experimental semantics" that has an entirely different understanding of the topic (see Matlock & Winter, 2015), coming more from the cognitive linguistic tradition, focusing on conceptualization, embodied cognition and all that fluffy stuff. The present introduction is FAR too jargony for a general audience, and I think it's even

too specific to semanticists/pragmaticists. If a semanticist comes from a different tradition (like myself), then terms such as "projectivity" or "at-issueness" are opaque as these relate to topics that are only discussed in the more formal tradition of semantics. (And some more malicious readers may throw the chapter angrily away because they feel that their understanding of semantics is not represented here) It's ok to focus on the experiment that the authors discuss (which is really useful as a case study), but explaining the concepts more for a general audience of linguists and acknowledging other traditions of semantics would be useful to draw more readers in.

We were under the impression that the readers of our DMUC would be formal semanticists/pragmaticists who want to start running their own experiments. We have now expanded the section that reports the content of the project to make it easier to read for non-specialists. We have also changed the framing somewhat to make the DMUC accessible to a broader audience of linguists who are interested in running web experiments.

2) The structure
The paper jumps straight into the specifics and then opens up to more general topics later. I'd reverse the structure, going from the most general to the more specific. First, the "summary of tools" is really not a summary but an instruction to clone the git repo so that the examples can be followed. There are many people who haven't even incorporated Git and MTurk into their work flow and some of these things are only explained later. These instructions need to come first. I'd also include a quick discussion of the utility of MTurk and web experiments to motivate the reader, citing some of the studies that have demonstrated the utility and data quality of MTurk. Then explain that you are going to introduce ONE way a web experiment work flow can be efficiently managed that you recommend for these and these reasons. At the moment, some of the motivation for the workflow actually come AFTER the work flow has been introduced (the reverse is better).

We have gotten rid of the “summary of tools” section and changed the structure and content of the paper as suggested, including going from the more general to the more specific, adding more motivation for and explanation of tools, and adding a discussion of the utility of web experiments.

The authors say on p. 11 that "we have used the same general infrastructure to run truth-value judgment studies, response time studies, self-paced reading studies, perception studies, free and forced production studies, studies involving drag-and-drop functionality, and many others."
-> This sort of information needs to come at the beginning.

We moved this to the beginning of the section.

The authors use EPrime later to mention how using proprietary software works against sharability. This is also reflected in a footnote on R/RStudio versus SPSS. I whole-heartedly agree but I think that this should also be frontloaded as it motivates everything else (and the R

footnote could be part of the main text). So what about signposting earlier "A common theme of this work flow is that it avoids any proprietary software. This increases the sharability ... and ultimately the reproducibility ... "?

We now say more clearly upfront in section 2 that the avoidance of proprietary software supports open science.

Jargony terms that need to be avoided:
- entailment-canceling operator
- at-issueness
- projection / projectivity / projective content

Rather than avoiding these terms, since they are foundational terms in the formal semantics literature on the topic, we have chosen to expand on them somewhat and hope that with the additional elaboration this section is easier to read. We also believe that the particular theoretical details are unimportant to the question of how best to organize an experimental project, and hope to have made clearer the generalizability of the framework to essentially any experimental project in both the abstract and Section 1.

Minor comments:
- p. 2, "We close with some reflections on what we would change, were we to start again from scratch." -> This is a bit oddly phrased. What about something like "Future directions"?

We rephrased this somewhat -- the point was to discuss some of the things we've learned in the meantime and what we would do differently now (eg with regards to pre-registration).

- p. 3, This is the first time Mechanical Turk is mentioned and it's already assumed. Many people actually don't know this. If the DMU cases are meant to be educational and tutorial-style, this needs to be explained.

We have added a little more explanation about Mechanical Turk at the beginning of section 3.

- p. 4, In my mind at least R markdowns are more reproducible than .R files — as they allow seeing both input and output. Just a thought that the authors COULD provide markdown files. Not a big issue though as I can also understand that the authors might want to keep the neat structure they have setup (very clean GitHub repo!).

At the time we didn't yet use Rmd. Given that this is a real use case, we don't want to report something we didn't do. We have, however, added some notes on the utility of using markdown in section 5.1, among the things we'd do differently today.

- p. 5, Footnote 5: "The broader project repository" -> "broad" seems off here; and perhaps just delete the "this repository is not as well documented for external eyes, but we're sharing it regardless", which sounds a bit odd.

Deleted in the reorganization of this section.

- p. 7+, The list of recommended project directories doesn't mention a "figures" folder although this seems clearly intended base don the Git repo structure. Also, I'd perhaps talk of a "figures" directory rather than "graphs" as one may also want to automatically generate figures that aren't data graphs (like visualizations of computational models etc.)

Our figures folders typically live inside the results folder (where we've called it "graphs") or inside the writing folder, if there are paper-specific figures to display. We try to avoid having a figures folder live on its own in the root directory, so we have not included it here.

- p. 11, Can we perhaps not be so negative of Qualtrics? I'm a fervent advocate of Qualtrics. I recognize that it's proprietary, but the speed with which sleek looking surveys (even with sophisticated randomization) can be generated is impeccable. I've also not found any limitations of Qualtrics for my research. I think it's cool to talk about the pro's and con's of different methods without immediately taking sides, which is also likely to throw less readers off.

We tried to tone down the rhetoric a little. Qualtrics makes running experiments at all an option for many, which we appreciate. But it remains true that coding up an experiment without being bound to Qualtrics allows for more flexibility. Eg, we can run interactive free production tasks or self-paced reading tasks by just coding them up as websites. We have opted for not discussing the pros and cons of different methods -- that would force us to discuss the other options besides Qualtrics that are out there when we only have limited space so we've decided to focus on what happened in this particular use case.

- p. 12, "with a participant's" -> error

Fixed.

- p. 13, "experiments for which were run before"

Fixed.

- p. 15, Isn't the lsmeans package surpassed by emmeans?

Indeed. We have now added a remark to this effect.

- p. 15, Footnote 6. I think there are more reasons that can be mentioned, such as 1) sharability, 2) R has more functionality thanks to its large developer base (many methods not implemented

in SPSS), 3) automation and 4) more freely available online documentation. Many of these things also pertain to using the open experiment workflow, so this could be in the section that argues against proprietary software.

Thanks, we've added these.

- p. 17, "and dreaming up exploratory analyses" -> I agree but this is FAR too accusatory in tone; this will throw some readers off

Indeed, modified.

- p. 17, In the discussion on preregistration, I'd mention in one or two sentences perhaps that preregistration != straitjacketing oneself, which is what many people believe. The preregistration CAN be changed but then at least changes are recorded and the distinction between confirmatory and exploratory analyses is public.

Thanks, we added some remarks to this effect.

Link to the chapter I mentioned:

http://bodowinter.com/stuff/Matlock&Winter_OxfordPress_ExperimentalSemantics.pdf

Matlock, T., & Winter, B. (2015). Experimental semantics. In B. Heine and H. Narrog (Eds.), Oxford Handbook of Linguistic Analysis (pp. 771-790). Oxford University Press.

Anyway, great stuff — was happy to read this contribution!

Recommendation: Revisions Required

Reviewer B:

Overall assessment

The paper does an excellent job of detailing the workflow of a specific project and attending to reproducibility. Recommend that the paper be accepted with very minor revisions required.

Minor issues

1. Per volume guidelines, the word 'speaker' on pg. 2 could be changed to 'language user' or something more neutral.

In this particular case, it is really the speaker/writer, given that we're dealing with a particular predicate of English.

2. For the list at the bottom of pg. 10 (Section 4.1), recommend that the dashes be changed to a numbered list to match the style used in Section 3.2, pgs. 6-8 (and other such lists, e.g., in Sections 4.2 & 5)

Fixed.

3. On pg. 12, there seem to be a few minor font issues - 'Figure 1' and the '3' for point three are in a different font than the rest of the text. These appear to be in Calibri, which the guidelines state should not be used.

Fixed.

4. Use tab indents for new paragraphs as per guidelines.

We have fixed this to the best of our ability by hand.

5. Spell out 'HIT' upon first use (pg. 9).

Fixed.

6. On pg. 9, the authors reference that participatn IDs are anonymized. Is there a record kept of participant IDs at all? If so, explain how this is handled within the workflow since people wanting to apply this framework will have to deal with this.

Yes, participant IDs are kept in the original files generated by supersubmiterator. We have now added a sentence explaining that the participant ID record is thus only kept on the researcher's machine who actually ran the experiment.

7. On pg. 17, the subsection should be 6.2, not 4.2

Fixed.

8. Same for 4.3/6.3. But, we recommend that 6.3 not be its own subsection, and that the content be folded into the opening paragraph of Section 6.

We have deleted that last subsection since it added nothing. We did, however, add a final section on how to adopt the overall framework and workflow, including Table 1 that includes an overview of all the tools used.

Other thoughts and recommendations

1. The mentioning of cognitive models in the abstract seems to indicate that more attention will be paid to this than actually is in Section 6.2. The authors may downplay this in the abstract a bit (and replace with more discussion of the key contributions); alternatively, the authors could be more specific in Section 6.2.

We have downplayed it in the abstract, since it's really not the main focus of the DMUC.

2. On pg. 4: "RStudio is also highly recommended." Agreed! But, could you add a very brief statement as to why, for those who may not already use it?

We have now added motivation.

3. Footnote one mentions that the referenced experiments will not work with IE. Are there any other dependencies that might become an issue over time in order to view these experiments, which might point to a broader issue of consideration in developing similar studies? (e.g., any dependencies on technologies that require ongoing support, which sometimes gets dropped [think of Java applets])

As long as html/javascript/css remain the foundation of web programming, this web experiment infrastructure will work.

4. On pg. 6 the authors refer to the git/github/Slack bundle as a "collaboration framework." This is a specific term, which some linguists might not be familiar with. It is really helpful - and this is a great point that the authors are making - so it may be good to offer a brief epansion as to what this is and why it is important.

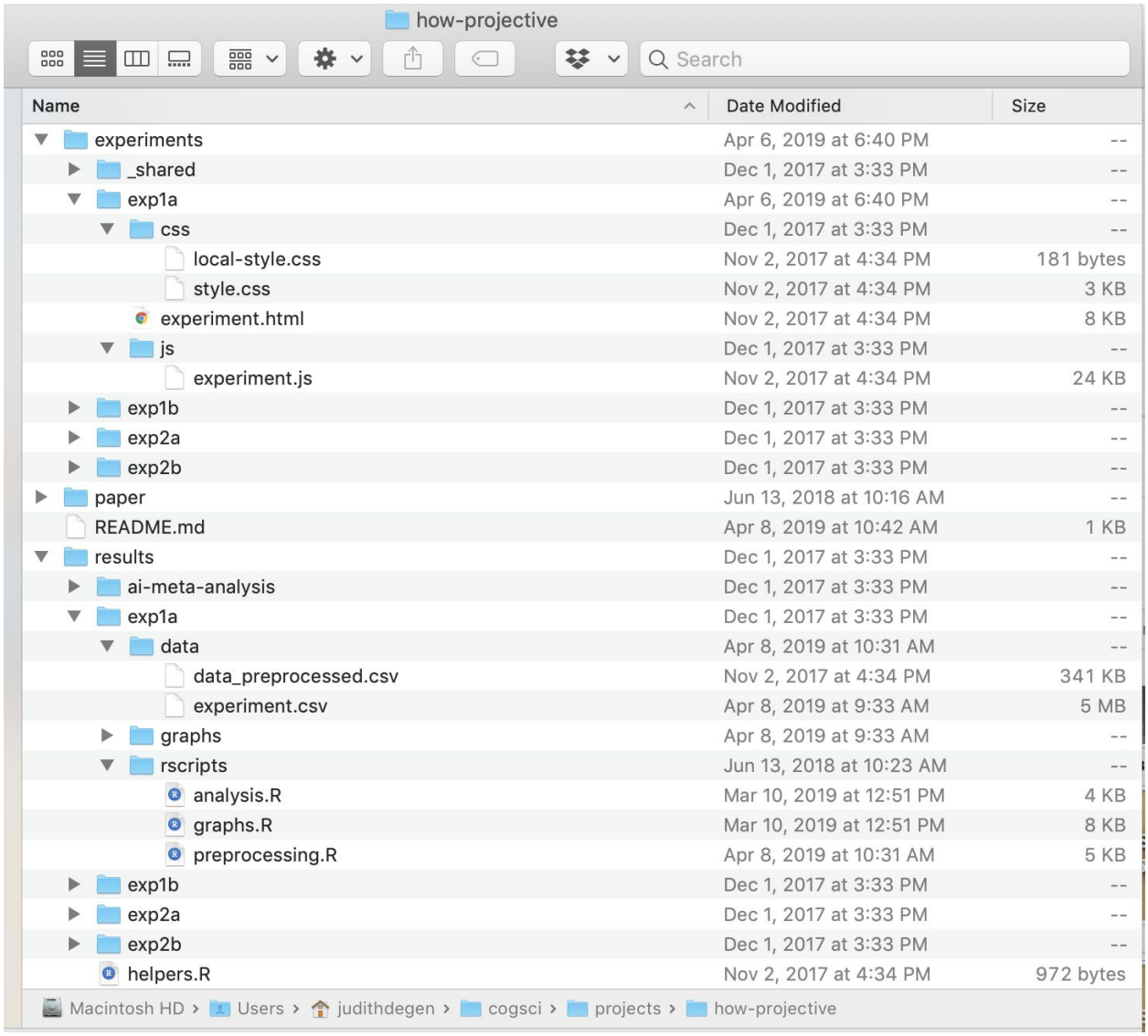
We have added prose in various places in section 3.1 that makes clearer that experimental work is often collaborative and therefore requires managing communication and file sharing.

5. Also on pg. 6 the authors introduce use of Slack. It may be helpful to add a statement about privacy issues related to this site. The authors do address privacy when talking about MTurk. It would be helpful to have here as well, especially if Slack is at all used to share files.

I'm not sure what you have in mind? Is the issue that Slack owns all the messages?

6. Images/screenshots might be a nice addition to visualize the project repository structure. A workflow visualization could also be good.

This is the most compact screenshot of the project directory we could get, and it is still very large and unwieldy:



We have therefore decided not to include the screenshot, given also that we provide 3 different ways in which the reader can access the directory directly, either in the browser or locally. But we are happy to include this screenshot if you think it adds value.

We have not included a visualization of the workflow, but instead included a table (Table 1) that contains an overview of all the tools that were used, and for what purpose. This should allow the

reader to more quickly extract the relevant information regarding which tool to use for which purpose.

7. In Section 4, a simple statement re: participant compensation might be nice (even in a footnote). MTurk has recently been getting attention because of the meager compensation of study participants.

We have added a remark explaining our own payment practices.

8. On pg. 13, line 3, suggest providing in the parenthetical statement a very brief reason as to why this is highly discouraged, just for clarity.

Added.

Recommendation: Revisions Required
