THE ROYAL SOCIETY PUBLISHING

PROCEEDINGS B

Larger communities create more systematic languages

Limor Raviv, Antje Meyer and Shiri Lev-Ari

Article citation details
Proc. R. Soc. B **286**: 20191262.
http://dx.doi.org/10.1098/rspb.2019.1262

Review timeline

Original submission:	11 February 2019
1st revised submission:	29 May 2019
2nd revised submission:	29 June 2019
Final acceptance:	1 July 2019

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2019-0145.R0 (Original submission)

Review form: Reviewer 1 (Dr Mark Atkinson)

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Reports © 2019 The Reviewers; Decision Letters © 2019 The Reviewers and Editors; Responses © 2019 The Reviewers, Editors and Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Please see attached file. (Appendix A)

Review form: Reviewer 2 (Matthew Lou-Magnuson)

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
The topic of community size and the role it plays in language structure is at present an open question. However, recent work has suggested that community size is likely conflated with other aspects of social network structure. For example, Lou-Magnuson & Onnis, 2018 has shown that the clustering coefficient predicts the capacity of a social network to support the development of linguistic complexity. In addition, authors cited in the paper address this issue of size as encapsulating more fundamental variables: Reali et al.(2018) state that in the social network data they gather from cellular phone communications, the clustering coefficients of the networks is a near constant factor, which they list a major limitation of the model.

From a linguistics prospective cited in the paper, Trudgill (2002) argues that intimacy of ties (as found in small communities and lost in larger communities) is the chief causal predictor of language complexity. Further Milroy & Milroy (1985) intimate that connection density is the driving force behind language innovation and simplification. Both works present support for the idea that dense connectivity, as found in smaller social networks, creates resistance to change (slowing innovation) but is required to develop greater levels of syntactic and morphological compositionality. In fact, Evans et al. (2009) suggests that not only the structural complexity is driven by small, intimate societies, but that it is needed to drive semantic innovations as well.

In light of such findings from both cognitive modeling and traditional linguistic approaches, I think this work could significantly contribute to the debate if it also addresses the role of connectivity in the social network, along side population size. Looking at the micro-networks presented in Figure 1, both networks are completely connected, thus having the same density/clustering coefficient, as identified in Reali et al. (2018) as problematic. While manipulation of connectivity is difficult to do in the small network of 4 participants (there are only 6 possible edges), the large network of 8 participants could be re-constructed with a lower degree of connectivity. I would like to see another condition, conducted on an 8 participant network, in which connection density or clustering coefficient are low. Reali et al. (2018) and Lou-Magnuson & Onnis (2018) both suggest values for theses measures as well differences in density vs. clustering that would be relevant.

Finally, as the paper address compositionality in human language, as opposed to animal communication, I would like to see the notion of compositionality treated in greater detail. Specifically, I would like to see mention of the difference between syntactic and morphological composition and the relation their experiment has for these two domains. For example, is the language game played by the participants capable of capturing this distinction? If not, how might it be altered to address the issue?

References
==========
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and brain sciences, 32(5), 429-448.

Lou-Magnuson, M., & Onnis, L. (2018). Social Network Limits Language Complexity. Cognitive science, 42(8), 2790-2817.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. Journal of linguistics, 21(2), 339-384.

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. Proceedings of the Royal Society B: Biological Sciences, 285(1871), 20172586.

Trudgill, P. (2002). Linguistic and Social Typology. In The Handbook of Language Variation and Change, J. K. Chambers, 481 P. Trudgill, and N. Schilling-Estes, eds. (Oxford, UK: Blackwell Publishing Ltd), pp. 707–728.

Decision letter (RSPB-2019-0145.R0)

14-Mar-2019

Dear Miss Raviv:

I am writing to inform you that your manuscript RSPB-2019-0145 entitled "Larger communities create more systematic languages" has, in its current form, been rejected for publication in Proceedings B.

This action has been taken on the advice of referees, who have recommended that substantial revisions are necessary. With this in mind we would be happy to consider a resubmission, provided the comments of the referees are fully addressed. However please note that this is not a provisional acceptance.

The resubmission will be treated as a new manuscript. However, we will approach the same reviewers if they are available and it is deemed appropriate to do so by the Editor. Please note that resubmissions must be submitted within six months of the date of this email. In exceptional circumstances, extensions may be possible if agreed with the Editorial Office. Manuscripts submitted after this date will be automatically rejected.

Please find below the comments made by the referees, not including confidential reports to the Editor, which I hope you will find useful. If you do choose to resubmit your manuscript, please upload the following:

1) A ‘response to referees’ document including details of how you have responded to the comments, and the adjustments you have made.
2) A clean copy of the manuscript and one with 'tracked changes' indicating your 'response to referees' comments document.
3) Line numbers in your main document.

To upload a resubmitted manuscript, log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Resubmission." Please be sure to indicate in your cover letter that it is a resubmission, and supply the previous reference number.

Sincerely,

Prof Sarah F. Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Comments to Author:
The two reviewers agree, as do I, that the paper addresses important and interesting issues. Reviewer 1 in particular is very positive, and provides only relatively minor suggestions to improve the clarity of the manuscript and situate the work more clearly in the context of the wider literature. However, Reviewer 2 is rather more critical. In particular, this reviewer raises the important point that the connectivity of social networks, rather than just absolute community size may be critical in the evolution of language structure. This parallels arguments in the comparative literature on the role of social factors in driving cognitive evolution, where researchers are increasingly moving away from crude metrics such as group size to look at more nuanced aspects of social complexity. If your data allow you to address this issue explicitly, I would strongly urge you to do so, as it could really help to improve the power of the study. If it is not possible to examine the issue empirically, then it should at least be considered carefully in the discussion.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)
Please see attached file.

Referee: 2

Comments to the Author(s)
The topic of community size and the role it plays in language structure is at present an open question. However, recent work has suggested that community size is likely conflated with other aspects of social network structure. For example, Lou-Magnuson & Onnis, 2018 has shown that the clustering coefficient predicts the capacity of a social network to support the development of linguistic complexity. In addition, authors cited in the paper address this issue of size as encapsulating more fundamental variables: Reali et al.(2018) state that in the social network data they gather from cellular phone communications, the clustering coefficients of the networks is a near constant factor, which they list a major limitation of the model.

From a linguistics prospective cited in the paper, Trudgill (2002) argues that intimacy of ties (as found in small communities and lost in larger communities) is the chief causal predictor of language complexity. Further Milroy & Milroy (1985) intimate that connection density is the driving force behind language innovation and simplification. Both works present support for the

idea that dense connectivity, as found in smaller social networks, creates resistance to change (slowing innovation) but is required to develop greater levels of syntactic and morphological compositionality. In fact, Evans et al. (2009) suggests that not only the structural complexity is driven by small, intimate societies, but that it is needed to drive semantic innovations as well.

In light of such findings from both cognitive modeling and traditional linguistic approaches, I think this work could significantly contribute to the debate if it also addresses the role of connectivity in the social network, along side population size. Looking at the micro-networks presented in Figure 1, both networks are completely connected, thus having the same density/clustering coefficient, as identified in Reali et al. (2018) as problematic. While manipulation of connectivity is difficult to do in the small network of 4 participants (there are only 6 possible edges), the large network of 8 participants could be re-constructed with a lower degree of connectivity. I would like to see another condition, conducted on an 8 participant network, in which connection density or clustering coefficient are low. Reali et al. (2018) and Lou-Magnuson & Onnis (2018) both suggest values for theses measures as well differences in density vs. clustering that would be relevant.

Finally, as the paper address compositionality in human language, as opposed to animal communication, I would like to see the notion of compositionality treated in greater detail. Specifically, I would like to see mention of the difference between syntactic and morphological composition and the relation their experiment has for these two domains. For example, is the language game played by the participants capable of capturing this distinction? If not, how might it be altered to address the issue?

References
==========

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and brain sciences, 32(5), 429-448.

Lou-Magnuson, M., & Onnis, L. (2018). Social Network Limits Language Complexity. Cognitive science, 42(8), 2790-2817.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. Journal of linguistics, 21(2), 339-384.

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. Proceedings of the Royal Society B: Biological Sciences, 285(1871), 20172586.

Trudgill, P. (2002). Linguistic and Social Typology. In The Handbook of Language Variation and Change, J. K. Chambers,
481 P. Trudgill, and N. Schilling-Estes, eds. (Oxford, UK: Blackwell Publishing Ltd), pp. 707–728.

Author's Response to Decision Letter for (RSPB-2019-0145.R0)

See Appendix B.

RSPB-2019-1262.R0

Review form: Reviewer 1 (Mark Atkinson)

Recommendation
Accept as is

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors have improved what was already a very good manuscript, and I recommend that it is accepted for publication without the need for further revision. All of my comments about the previous version of the paper have been very thoroughly addressed, and I am entirely satisfied with the responses.

Decision letter (RSPB-2019-1262.R0)

25-Jun-2019

Dear Miss Raviv

I am pleased to inform you that your manuscript RSPB-2019-1262 entitled "Larger communities create more systematic languages" has been accepted for publication in Proceedings B.

If you have any last changes on your manuscript, log into https://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision. You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre. Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days. If you do not think you will be able to meet this date please let us know.

Before uploading your revised files please make sure that you have:

1) A text file of the manuscript (doc, txt, rtf or tex), including the references, tables (including captions) and figure captions. Please remove any tracked changes from the text before submission. PDF files are not an accepted format for the "Main Document".

2) A separate electronic file of each figure (tiff, EPS or print-quality PDF preferred). The format should be produced directly from original creation package, or original software format. PowerPoint files are not accepted.

3) Electronic supplementary material: this should be contained in a separate file and where possible, all ESM should be combined into a single file. All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

4) A media summary: a short non-technical summary (up to 100 words) of the key findings/importance of your manuscript.

5) Data accessibility section and data citation
It is a condition of publication that data supporting your paper are made available either in the electronic supplementary material or through an appropriate repository.

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should be fully cited. To ensure archived data are available to readers, authors should include a ‘data accessibility’ section immediately after the acknowledgements section.

This should list the database and accession number for all data from the article that has been made publicly available, for instance:
• DNA sequences: Genbank accessions F234391-F234402
• Phylogenetic data: TreeBASE accession number S9123
• Final DNA sequence assembly uploaded as online supplemental material
• Climate data and MaxEnt input files: Dryad doi:10.5521/dryad.12311
NB. From April 1 2013, peer reviewed articles based on research funded wholly or partly by RCUK must include, if applicable, a statement on how the underlying research materials – such as data, samples or models – can be accessed. This statement should be included in the data accessibility section.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available) which will take you to your unique entry in the Dryad repository. If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link. Please see https://royalsociety.org/journals/ethics-policies/data-sharing-mining/ for more details.

6) For more information on our Licence to Publish, Open Access, Cover images and Media summaries, please visit https://royalsociety.org/journals/authors/author-guidelines/.

Once again, thank you for submitting your manuscript to Proceedings B and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Sincerely,

Dr Sarah Brosnan
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member
Comments to Author:
I found the revisions to be thorough and comprehensive, and the paper is much improved as a result. The authors have clearly addressed the points raised by myself and the reviewers, and the paper makes an excellent contribution to the literature.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s).
The authors have improved what was already a very good manuscript, and I recommend that it is accepted for publication without the need for further revision. All of my comments about the previous version of the paper have been very thoroughly addressed, and I am entirely satisfied with the responses.

Decision letter (RSPB-2019-1262.R1)

01-Jul-2019

Dear Miss Raviv

I am pleased to inform you that your manuscript entitled "Larger communities create more systematic languages" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Paper charges
An e-mail request for payment of any related charges will be sent out shortly. The preferred payment method is by credit card; however, other payment options are available.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Appendix A

Raviv et al. assess the proposal that more systematic languages are more likely to emerge in larger communities, and find experimental evidence in support of it. The study is well-designed, well-motivated by the literature, and the conclusions are clear and supported by the data. I recommend that it is accepted for publication in Proceedings B and anticipate that it would be of interest to a more general readership, as well as to researchers directly involved in answering similar research questions.

Though I would be broadly happy to see the paper accepted as it is, I do make a number of suggestions below which I think would improve it. I stress that these are all relatively minor points and I would anticipate that the authors would be able to address them without difficultly.

1. There is other published experimental work which I think it would be good to mention to see how this study fits in with the rest of the literature. I honestly don't want to be a reviewer who asks for their own papers to be cited for the sake of it, but in each of Atkinson, Kirby, and Smith (2015), Atkinson, Smith, and Kirby (2018), and Atkinson, Mills, and Smith (2018), we consider the role of group size and input variability and language complexity, and find no evidence for a direct influence of population size on language complexity. Each of these studies considers how community size may influence language complexity in a different way to how it's been done in this study, and so there's no suggestion of one set of results invalidating another as far as I'm concerned. But (and I know that the first author will be aware than Kenny Smith has already made this point elsewhere), it will be by considering all experimental work on the effects of community size on language features that we will learn about the specific elements of language influenced by community size and the mechanisms by which community size has such effects.

2. In a similar vein to the previous point and helping see how this result fits into the literature, I think it's also worth at least briefly discussing how these results may, or may not, relate to community size effects in real languages, even if it's just acknowledging the limitations of the study and highlighting directions for future research. Here, you have only compared two, relatively small, group sizes, assuming a limited shared holistic lexicon at the start, and no population turnover. This is not a criticism of this study: lab-based experiments like this are valuable and often necessarily rely on relatively small group sizes and meaning spaces, and of course research like this has to start somewhere. But what would you expect to see, e.g., if this experiment was scaled up to consider more of a difference between the group sizes? Or if there was also language learning involved as new individuals entered a group, given that population turnover may also lead to an increase in systematicity?

3. The interpretation of p-values between 0.05 and 0.1 in incorrect. A p-value of 0.078 (line 306), assuming a 5% significance threshold, provides no evidence for an effect of group size; it does not imply that the dependent variable was "marginally modulated". See also lines 308 and 349.

4. p.2 para 2 and final paragraph of the discussion: Should the references to Everett et al. (12) not be to Lupyan and Dale (1)?

5. p.2 line 58: "be more geographically spread"?

6. p.2 para 3: You've referenced Wray and Grace for larger communities having a larger proportion of non-native speakers etc., but I think they just propose that rather than put forward any hard evidence for it. I'd suggest making that clear and/or citing Lupyan and Dale or Bentz and Winter, who at least use Ethnologue data to support the link between population size and proportion of adult learners.

7. p.2 para 4: I'd clarify what you mean by "input variability" here, as the context is a little different to how I think it is usually used, i.e. in the unidirectional receiving of data by a learner (as in, e.g., Atkinson, Smith, and Kirby, 2018). The type of input you're talking about is also quite different from that of the studies you cite at the end of the paragraph.

8. line 157-8: "To establish common grounds" wasn't clear to me here.

9. line 186-7: Can you make it clear whether the data for these groups was excluded or not. On first reading, I assumed the data for these groups was discarded from the analysis.

10. c. lines 220-224: I accept the difficulty with using z-scores here, but can you justify why using the raw correlation scores as opposed to z-scores isn't problematic here, other than citing your previous work as a precedent? One option may be to use z-scores on a subset of your analysis, e.g. just the Round 16 data or the data for the models of type (II) to confirm that it doesn't change your interpretation of the data?

11. p.8 and elsewhere: Can you motivate why the models of type (II) specifically compared the data at Rounds 15-16? Is there not something a bit weird about pooling data arising from communication and individual testing in this way? That said, re-running your analysis on different subsets of your data, e.g. just the Round 16 data, doesn't seem to affect the pattern of results.

12. p.8: this'll probably be picked up later, but there are a number of typos on this page: "calcualted", "avarge", "intrecpects".

13. In at least one case, one of the reported models indicates a singular fit: reg_str_new_p in the Rmd file. Alternative models do not change the general pattern of results as far as I can tell, but it might be worth checking the models and updating the model outputs accordingly.

14. Appendix C reference to Becker should be (46), not (47). And maybe add a note that the first page or two of text is the same as in the main paper (or don't replicate it?)?

References

Atkinson, M., Mills, G. J., Smith, K. 2018. Social group effects on the emergence of communicative conventions and language complexity. Journal of Language Evolution. https://doi.org/10.1093/jole/lzy010.

Atkinson, M., Smith, K., Kirby, S. 2018. Adult learning and language simplification. Cognitive Science, 42, 2818-2854.

Atkinson, M., Kirby, S., Smith, K., 2015. Speaker input variability does not explain why larger populations have simpler languages. PLOS ONE, 10(6): e0129463.

Appendix B

29th of May, 2019

Dear Prof. Brosnan,

Many thanks to you and the reviewers for your detailed and insightful comments. We have read them very carefully, and believe we have successfully addressed all of the concerns in our revision.

Both reviewers felt that the paper addresses an interesting and important question in the field, and that it presents novel empirical findings that are of interest to the wide research community. Yet they raised several theoretical and empirical issues, all of which have been addressed in the revision, and are referred to in detail below.

The main concern, which was raised by Reviewer 2 and mentioned by the Editor, had to do with the hypothesis that network structure and network connectivity, which are typically confounded with community size in the real-world, play a more prominent role in explaining cross-linguistic differences in structural complexity compared to community size. Specifically, theories of language change suggest that differences in network density may be the true underling mechanism behind language simplification (Trudgill, 2002; Milroy & Milroy, 1985). This idea is supported by computational work showing that differences in the structural properties of networks (e.g., their degree of clustering and hierarchy) can lead to differences in linguistic complexity and compositionality, and that these effects are further modulated by population size (Lou- Magnuson & Onnis, 2018). Together, it implies that network structure may be a highly important social feature in predicting patterns of language diversity, alongside (and perhaps even more than) community size.

We completely agree that examining the role of network structure is of high interest, and we share the view that it should be experimentally tested and teased apart from that of community size, especially due to its conflating potential. We also mention this issue in our introduction (page 2). However, we think that different social features should be tested **separately** in order to examine their unique and causal contribution. As such, the goal of the current paper was to single out and specifically tease apart one of these confounding features (namely, community size), while controlling for the others. Accordingly, we chose to manipulate only the number of group members, while keeping the groups' structure constant (i.e., fully connected network). Although the effects of network structure and

community size may be non-additive, understanding those interactions requires that we first understand the individual effect of each feature. Our results show that when network structure is kept constant, differences in community size alone can affect the emergence of linguistic structure, and shed light on the possible reasons why larger groups develop more compositionality over time. The current study therefore suggests that community size has its own unique role in language emergence, above and beyond network structure, and attempts to explain one of its underlying mechanisms (i.e., input variability). We believe these results are important and rich enough to justify a separate publication.

While network structure could potentially be manipulated in addition to community size (i.e., by introducing a third condition, as suggested by the Reviewer), we believe that properly discussing and examining the effect of network structure requires (a) addressing a different, and far too large, body of literature, as well as (b) testing different types of network structure conditions that address different aspects of social networks (e.g., network hierarchy and clustering). Sadly, the scope and length restrictions of the current paper do not allow us to do this in the necessary depth. It would not be possible to include the needed additional conditions, analyses and discussions on the role of network structure without substantially cutting down on the current analyses and discussions, greatly hindering the depth and strength of the paper, which is already quite compact. Therefore, this paper focuses only on the property of community size, and provides the first comprehensive experimental examination of its individual role and its underlying mechanism. It is an important first step in experimentally teasing apart different social features that are confounded in the real-world, and establishes community size as one of the relevant properties in explaining patterns of language diversity. While we agree that network structure may be just as relevant (or even more), we believe it should be studied independently.

We are currently running a series of experiments to test the individual role of network structure, and their design is quite similar to that the reviewer suggests. We are using the same paradigm as in the current study to contrast groups of the same size (eight participants) in three different network conditions: (i) fully-connected networks, (ii) small-world networks, and (iii) scale-free networks. To our surprise, preliminary results from the groups collected so far show no difference between the three types of networks. Across conditions, we see similar trajectories of linguistic structure emergence, with all networks reaching the same levels of compositionality by the end of the experiment. This null result extended to the other measures (e.g., stability, convergence) as well. While we cannot draw

strong conclusions from these results, they suggest that network structure may play a less prominent role in language emergence compared to community size (at least in a relatively small community). This idea is in line with the modelling work of Spike (2017), who found that as long as networks have small-world properties, their specific network structure has a relatively small role to play in the development and maintenance of linguistic complexity. We provide a more detailed explanation of the experimental setup and our findings so far (including relevant figures) in our response to Reviewer 2.

To summarize, we agree that network structure is an important feature to study, and we are currently attempting to do so. At the same time, we are of the opinion that it is outside the scope of the current paper, and our results so far suggest that it may even be of lesser importance to community size, which is the focus of this paper.

The reviewers raised several additional issues, all of which have been addressed in the revision. We will now outline the changes we have made in detail, in the order of the reviews' comments. We believe that we were able to address all the concerns and questions raised by the reviewers in the revised version, and we would like thank them for helping us improve the paper.

Sincerely,
The authors

Associate Editor:

The two reviewers agree, as do I, that the paper addresses important and interesting issues. Reviewer 1 in particular is very positive, and provides only relatively minor suggestions to improve the clarity of the manuscript and situate the work more clearly in the context of the wider literature. However, Reviewer 2 is rather more critical. In particular, this reviewer raises the important point that the connectivity of social networks, rather than just absolute community size may be critical in the evolution of language structure. This parallels arguments in the comparative literature on the role of social factors in driving cognitive evolution, where researchers are increasingly moving away from crude metrics such as group size to look at more nuanced aspects of social complexity.

If your data allow you to address this issue explicitly, I would strongly urge you to do so, as it could really help to improve the power of the study. If it is not possible to examine the issue empirically, then it should at least be considered carefully in the discussion.

- Given that we deliberately kept network structure similar across conditions, we are not able to directly test the role of network structure in the current data. Following the Editor's suggestion, we now clarify this issue in our revised introduction (page 2), and specifically mention in the Methods section that network structure was kept constant and fully-connected (page 3). We also get back to the potential role of network structure in our revised discussion, where we discuss the necessity of looking at other social features beyond community size, and specifically network structure (page 13).
- Given the journal's length restrictions, we were not able to discuss this issue in greater depth. Nevertheless, we believe we have now sufficiently acknowledged and clarified the idea that network structure is an additional relevant feature that can affect the emergence of linguistic structure, and that it should be teased apart from community size.

Reviewer 1:

Raviv et al. assess the proposal that more systematic languages are more likely to emerge in larger communities, and find experimental evidence in support of it. The study is well-designed, well-motivated by the literature, and the conclusions are clear and supported by the data. I recommend that it is accepted for publication in Proceedings B and anticipate that it would be of interest to a more general readership, as well as to researchers directly involved in answering similar research questions.

Though I would be broadly happy to see the paper accepted as it is, I do make a number of suggestions below which I think would improve it. I stress that these are all relatively minor points and I would anticipate that the authors would be able to address them without difficultly.

1. There is other published experimental work which I think it would be good to mention to see how this study fits in with the rest of the literature. I honestly don't want to be a reviewer who asks for their own papers to be cited for the sake of it, but in each of Atkinson, Kirby, and Smith (2015), Atkinson, Smith, and Kirby (2018), and Atkinson, Mills, and Smith (2018), we consider the role of group size and input variability and language complexity, and find no evidence for a direct influence of population size on language complexity. Each of these studies considers how community size may influence language complexity in a different way to how it's been done in this study, and so there's no suggestion of one set of results invalidating another as far as I'm concerned. But (and I know that the first author will be aware than Kenny Smith has already made this point elsewhere), it will be by considering all experimental work on the effects of community size on language features that we will learn about the specific elements of language influenced by community size and the mechanisms by which community size has such effects.

    - We thank the reviewer for the suggested citations, and we now report them in our revised introduction (page 2-3).
    - We would like to note that we did not mean to ignore this literature, and we apologize if that was the given impression. When we wrote the paper, the only published study we were aware of was Atkinson et al. (2015). Because we had a tight word limit to adhere to, and because that paper differs from our paper in the manner that it manipulates variability and the phenomenon it focuses on, we ended

up omitting its report. Following the reviewer's comment, we have now included all the suggested papers in our introduction.

2. In a similar vein to the previous point and helping see how this result fits into the literature, I think it's also worth at least briefly discussing how these results may, or may not, relate to community size effects in real languages, even if it's just acknowledging the limitations of the study and highlighting directions for future research. Here, you have only compared two, relatively small, group sizes, assuming a limited shared holistic lexicon at the start, and no population turnover. This is not a criticism of this study: lab-based experiments like this are valuable and often necessarily rely on relatively small group sizes and meaning spaces, and of course research like this has to start somewhere. But what would you expect to see, e.g., if this experiment was scaled up to consider more of a difference between the group sizes? Or if there was also language learning involved as new individuals entered a group, given that population turnover may also lead to an increase in systematicity?

   - We thank the reviewer for this comment, and we believe our lab-based design does scale up to larger and more realistic scenarios, if one considers that the amount of experience and familiarity with the language and with the other members of the community also scale up accordingly. That is, the meaning space and the length of the experiment should be proportionate to the population size. We now discuss this point in the revised manuscript (page 12). In our experiment, people interact with each other over a miniature meaning space for only a few hours, and never meet their partners more than a handful of times. In the real world, people interact with many more partners about many more meanings, but also have years and years to do so. In the current scaling ratios, doubling the number of participants in each group was a sufficiently strong manipulation that led to a significant difference in the trajectory of structure emergence. Therefore, we expect that as long as the experimental features are scaled accordingly, similar results should be obtained for scenarios where groups differ in their size even more extremely. In all such cases, larger groups are still faced with a greater communicative challenge where it is harder for them to converge on a shared lexicon compared to a smaller group, and are therefore under a stronger pressure to develop more systematic languages.
   - Given the current literature on iterated learning, it is highly likely that adding language learning by novices will introduce an additional pressure for systematicity, leading to more linguistic structure overall. Therefore, we predict

that introducing generation turnover (where new members enter the group) would amplify the effect of group size even more as long as the ratio of generation turnover is somehow proportionate to the community size, e.g., that larger groups have more individuals entering the community, or have more people replaced at each generation.

3. The interpretation of p-values between 0.05 and 0.1 in incorrect. A p-value of 0.078 (line 306), assuming a 5% significance threshold, provides no evidence for an effect of group size; it does not imply that the dependent variable was "marginally modulated". See also lines 308 and 349.

    - We thank the reviewer for this valid point, and we have removed all such descriptions from the revised manuscript.

4. p.2 para 2 and final paragraph of the discussion: Should the references to Everett et al. (12) not be to Lupyan and Dale (1)?

    - We thank the reviewer for noticing this mix-up, and we have now fixed all references numbers throughout the manuscript.

5. p.2 line 58: "be more geographically spread"?

    - With “more geographically spread” we meant that larger communities tend to occupy larger areas and spread out on a larger territories. We have now changed the phrasing to “geographically spread out” to help clarify this matter (page 2).

6. p.2 para 3: You've referenced Wray and Grace for larger communities having a larger proportion of non-native speakers etc., but I think they just propose that rather than put forward any hard evidence for it. I'd suggest making that clear and/or citing Lupyan and Dale or Bentz and Winter, who at least use Ethnologue data to support the link between population size and proportion of adult learners.

    - We thank the reviewer for this comment, and we have updated the references as suggested.

7. p.2 para 4: I'd clarify what you mean by "input variability" here, as the context is a little different to how I think it is usually used, i.e. in the unidirectional receiving of data by a learner (as in, e.g., Atkinson, Smith, and Kirby, 2018). The type of input you're talking about is also quite different from that of the studies you cite at the end of the paragraph.

    - The term "input variability" refers to the extent to which the available data points differ from each other, or in other words, the degree of variability in the information people are exposed to. This broad definition is in line with the input variability studies cited in the paper, as well as with Atkinson et al. (2018). Notably, psycholinguistic and culture evolution studies often use the number of models people can learn from (unidirectionally), and treat it as a proxy for the amount of variability in the input. While this is fairly reasonable, it is important to distinguish between this rich term and one of its practical implementations. We adopt the reviewer's suggestion and now clarify that by input variability we are referring to the number of different variants speakers are exposed to (page 2).

8. line 157-8: "To establish common grounds" wasn't clear to me here.

- With "to establish common grounds" we meant that participants would have a shared starting point. We have now changed the sentence to help clarify this idea (page 5).

9. line 186-7: Can you make it clear whether the data for these groups was excluded or not. On first reading, I assumed the data for these groups was discarded from the analysis.

    - The existing data from these two groups was included in the analyses. We have added a sentence to the text clarifying this issue (page 5).

10. lines 220-224: I accept the difficulty with using z-scores here, but can you justify why using the raw correlation scores as opposed to z-scores isn't problematic here, other than citing your previous work as a precedent? One option may be to use z-scores on a subset of your analysis, e.g. just the Round 16 data or the data for the models of type (II) to confirm that it doesn't change your interpretation of the data?

    - The decision to use the raw correlation scores (*r*) instead of the z-scores was based on careful examination of the properties of these measures, and was made in close

consultation with experts in the field (e.g., James Winters, Kevin Stadler & Matt Spike). In short, the use of z-scores to indicate an increase in compositionality can be mathematically problematic and misleading for various reasons, and especially since z-scores are inflated for larger data sets. We now clarify this issue in the Methods section (page 6).

- In general, z-scores test whether there is compositionality. They are not a measure of effect size. Our goal in the study, however, is not to test whether there is compositionality (there is in both conditions), but whether the degree of compositionality differs across the two conditions. Correlation scores reflect the magnitude of the compositionality more directly. Similar arguments can be found in detail in Stadler & Spike (in preparation): Measures of compositionality for artificial language experiments (https://github.com/kevinstadler/compositionality/blob/master/mantel-r-vs-z.pdf), who closely examined the use of z-scores and R-values using re-analyses of data obtained from iterated learning experiments. They conclude that: "*It should be stressed that the z-score is related to the significance level of a measure, rather than expressing the effect size of the measure itself. High z-scores therefore do not necessarily capture high levels of compositionality, rather than high (statistical) confidence in the presence of some level of compositionality (Spike, 2016, p.186)… The z-score really captures the significance level of the correlation, rather than the actual structure preserved by the signals' mapping between the form and meaning spaces. As a consequence, differences in the size of the test sets obtained by different experimental designs or conditions – or even by testing differences between generations – can therefore have an adverse effect on the interpretability of the measure (Cornish et al., 2009)…This effect is expected (and desired) for a measure of significance, but not indicative of an actual increase in compositionality. This suggests that* ***the z-score should be categorically avoided for drawing comparisons between compositionality levels of different-sized data sets".***
- As for the Reviewer's suggestion to only use z-scores on a subset of the data (e.g., Round 16) to ensure that the it doesn't change the interpretation of the data, it has been shown that z-scores and raw correlations follow each other in equally-sized data sets, making the additional measure redundant (See Figure 2 below from Stadler & Spike, showing how the two measures track each other almost perfectly). That is, rerunning the analyses using z-scores instead of raw correlations would yield the exact same results.

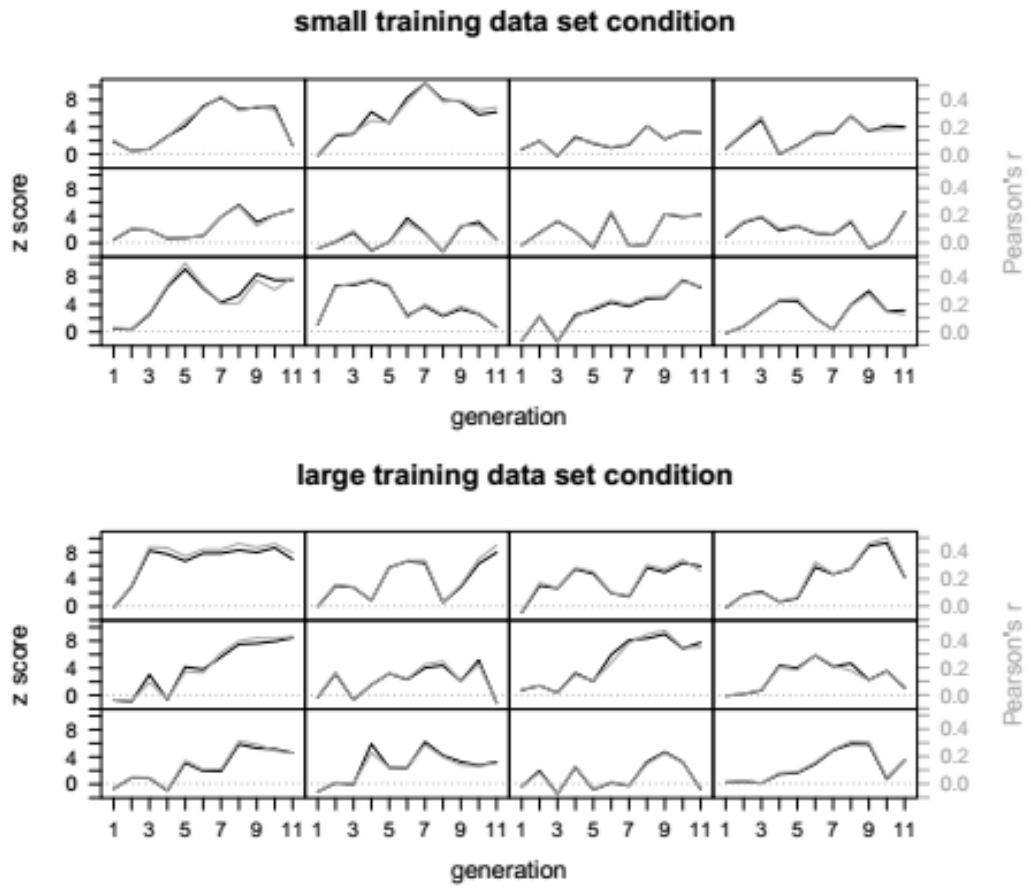
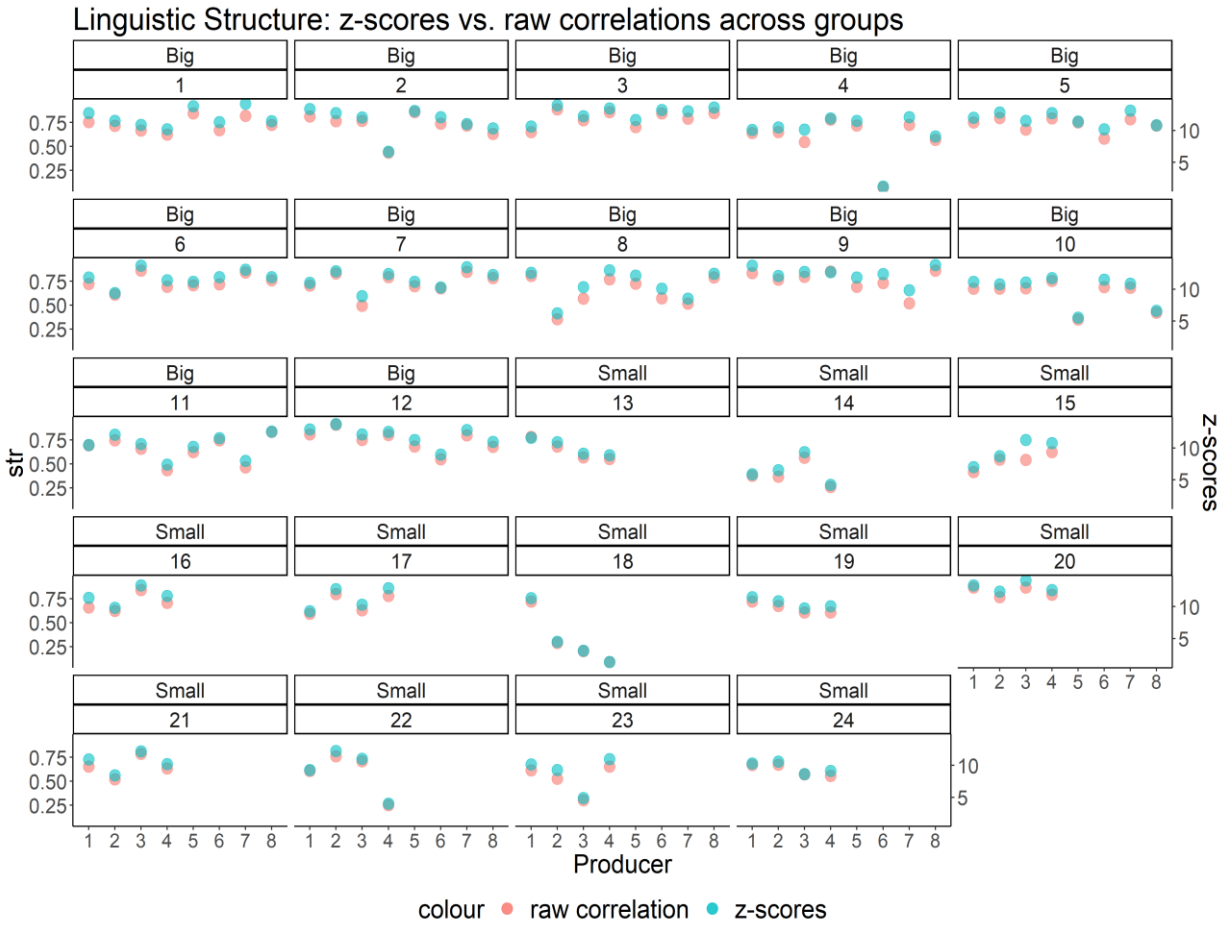


Figure 2: Comparison of the z score as derived from the distribution of rs for 1000 random permutations (black line, left axis) plotted against the raw value of Pearson's r (gray line, right axis) for the 24 iterated learning chains run by Beckner et al. (2017). The dotted gray line indicates the baseline of the correlation coefficient ($r = 0$) where there is no evidence for correlation (either positive or negative).

- Nevertheless, to address the Reviewer's concern we also rerun the Mantel test on the languages created by participants in Round 16 (as suggested by the Reviewer), and rerun the type II model with those z-scores as the dependent variable. This model was identical to the reported model that examined linguistic structure of the final languages (model #11) but without the nested random effect for participant given that the model only include round 16, and therefore each participant only had one score. The results confirm the pattern reported in the paper, and show that the final languages created by members of larger groups have significantly higher z-scores (see models for z-scores and raw correlation below). We also include a plot of the z-scores obtained from the Mantel test (in blue) alongside the original raw correlation scores used in the paper (in red) on the y-axes, for each participant (1 to 8, on the x-axis) across all 24 groups. As can be seen, the two measures align in our data as well.




```
> summary(reg_str_comp_end_p2)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: str ~ Condition + (1 | Group)
   Data: str_comp_end_p2

     AIC      BIC   logLik deviance df.resid
  -140.7   -128.8     74.3   -148.7      140

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.1412 -0.4016  0.1967  0.6192  2.1237

Random effects:
 Groups   Name        Variance Std.Dev.
 Group    (Intercept) 0.005932 0.07702
 Residual             0.017390 0.13187
Number of obs: 144, groups:  Group, 24

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.59561    0.02937  20.280
Condition1   0.10681    0.03914   2.729

Correlation of Fixed Effects:
           (Intr)
Condition1 -0.748
```

```
> summary(reg_str_comp_end_z)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: z ~ Condition + (1 | Group)
   Data: str_comp_end_p2

     AIC      BIC   logLik deviance df.resid
   647.1    658.9   -319.5    639.1      140

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.3390 -0.3981  0.1976  0.6468  2.1847

Random effects:
 Groups   Name        Variance Std.Dev.
 Group    (Intercept) 1.529    1.237
 Residual             4.090    2.022
Number of obs: 144, groups:  Group, 24

Fixed effects:
            Estimate Std. Error t value
(Intercept)   9.6402     0.4611  20.906
Condition1    1.6930     0.6186   2.737

Correlation of Fixed Effects:
            (Intr)
Condition1 -0.745
```

11. p.8 and elsewhere: Can you motivate why the models of type (II) specifically compared the data at Rounds 15-16? Is there not something a bit weird about pooling data arising from communication and individual testing in this way? That said, re-running your analysis on different subsets of your data, e.g. just the Round 16 data, doesn't seem to affect the pattern of results.

- We thank the reviewer for this comment, and for confirming our reported pattern of results in the subset analysis. We decided to run the models of combined data from round 15 (the last communication round) and round 16 (the last test round) in order to examine participants' final languages, and get a clear and comprehensive picture of the languages at their final state. Given the instructions of the experiment and its communicative goal, we do not believe that participants in the test phase would use different languages than those they used in the communication phase. Our decision to include the last two rounds of the experiment (rather than just the final test round) regardless of their type was done in order to (a) ensure there is sufficient data for the model to converge, and (b) ensure that our results are not driven by different behavior in communication vs. individual test. As the reviewer noted, running the analyses on only one of these rounds does not change the results.

12. p.8: this'll probably be picked up later, but there are a number of typos on this page: "calcualted", "avarge", "intrecpects".

- We thank the reviewer for noticing these typos. We have fixed them in the revised version.

13. In at least one case, one of the reported models indicates a singular fit: reg_str_new_p in the Rmd file. Alternative models do not change the general pattern of results as far as I can tell, but it might be worth checking the models and updating the model outputs accordingly.

- We double-checked and reran all models in the Rmd file, yet we did not find any model with a singular fit, including the suggested model reg_str_new_p. We attached a screen shot of this model's full summary below. That is, in our versions of R (3.3.3) and lme4 (1.1-18-1), we did not get any such issues. While it is possible that this issue arises only in another version of the program, we would like to reassure the reviewer that it is not the case for any of the models reported in the paper.

```
> summary(reg_str_new_p)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Structure ~ poly(c.Round, 2) * Condition + (1 + poly(c.Round,      2) | Group/Producer)
   Data: str_long_p

     AIC      BIC   logLik deviance df.resid
 -1955.1  -1846.0    996.6  -1993.1     2288

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.5040 -0.5766  0.0830  0.6268  2.6945

Random effects:
 Groups         Name                Variance Std.Dev. Corr
 Producer:Group (Intercept)         0.007029 0.08384
                poly(c.Round, 2)1 5.087261 2.25550  -0.30
                poly(c.Round, 2)2 2.416921 1.55465  -0.28 -0.66
 Group          (Intercept)         0.006619 0.08136
                poly(c.Round, 2)1 1.023344 1.01160   0.38
                poly(c.Round, 2)2 0.467235 0.68355  -0.72 -0.92
 Residual                           0.019507 0.13967
Number of obs: 2307, groups:  Producer:Group, 168; Group, 24

Fixed effects:
                            Estimate Std. Error t value
(Intercept)                  0.50149    0.02678  18.727
poly(c.Round, 2)1            4.55082    0.48083   9.465
poly(c.Round, 2)2           -3.00237    0.37622  -7.980
Condition1                   0.05621    0.03678   1.528
poly(c.Round, 2)1:Condition1 1.92134    0.62842   3.057
poly(c.Round, 2)2:Condition1 0.44736    0.48480   0.923

Correlation of Fixed Effects:
            (Intr) pl(.R,2)1 pl(.R,2)2 Cndtn1 p(.R,2)1:
ply(c.R,2)1  0.136
ply(c.R,2)2 -0.428 -0.500
Condition1  -0.728 -0.099     0.311
p(.R,2)1:C1 -0.104 -0.765     0.383     0.167
p(.R,2)2:C1  0.332  0.388    -0.776    -0.452 -0.542
>
```

14. Appendix C reference to Becker should be (46), not (47). And maybe add a note that the first page or two of text is the same as in the main paper (or don't replicate it?)?

- We thank the reviewer for this suggestion, and have accordingly removed the redundant text from Appendix C.

Reviewer 2:

The topic of community size and the role it plays in language structure is at present an open question. However, recent work has suggested that community size is likely conflated with other aspects of social network structure. For example, Lou-Magnuson & Onnis, 2018 has shown that the clustering coefficient predicts the capacity of a social network to support the development of linguistic complexity. In addition, authors cited in the paper address this issue of size as encapsulating more fundamental variables: Reali et al.(2018) state that in the social network data they gather from cellular phone communications, the clustering coefficients of the networks is a near constant factor, which they list a major limitation of the model.

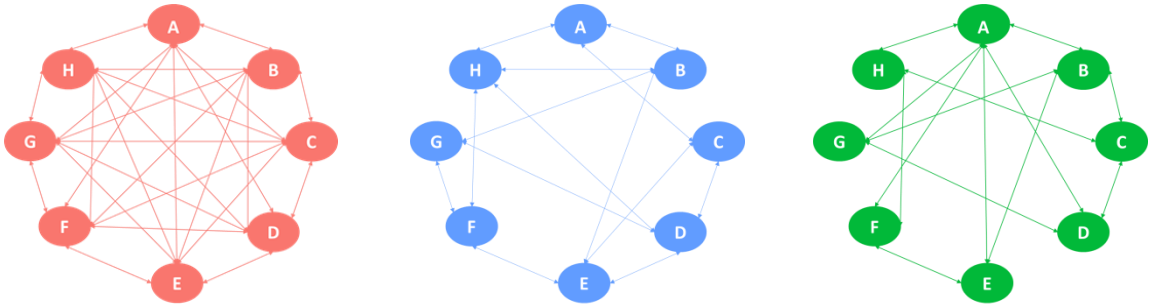
From a linguistics prospective cited in the paper, Trudgill (2002) argues that intimacy of ties (as found in small communities and lost in larger communities) is the chief causal predictor of language complexity. Further Milroy & Milroy (1985) intimate that connection density is the driving force behind language innovation and simplification. Both works present support for the idea that dense connectivity, as found in smaller social networks, creates resistance to change (slowing innovation) but is required to develop greater levels of syntactic and morphological compositionality. In fact, Evans et al. (2009) suggests that not only the structural complexity is driven by small, intimate societies, but that it is needed to drive semantic innovations as well.

In light of such findings from both cognitive modeling and traditional linguistic approaches, I think this work could significantly contribute to the debate if it also addresses the role of connectivity in the social network, alongside population size. Looking at the micro-networks presented in Figure 1, both networks are completely connected, thus having the same density/clustering coefficient, as identified in Reali et al. (2018) as problematic. While manipulation of connectivity is difficult to do in the small network of 4 participants (there are only 6 possible edges), the large network of 8 participants could

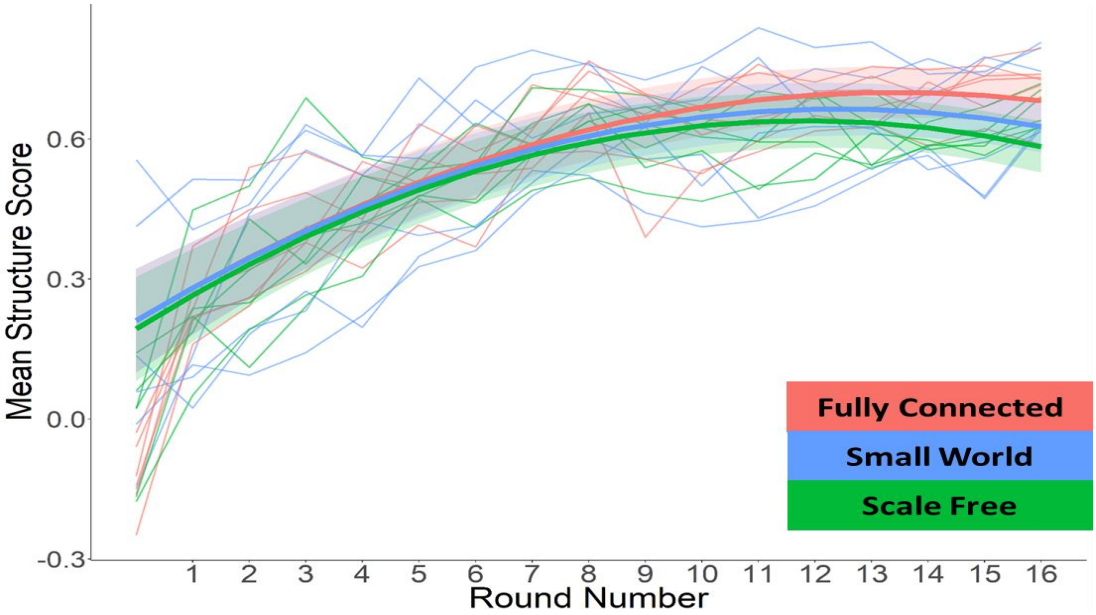
be re-constructed with a lower degree of connectivity. I would like to see another condition, conducted on an 8 participant network, in which connection density or clustering coefficient are low. Reali et al. (2018) and Lou-Magnuson & Onnis (2018) both suggest values for these measures as well differences in density vs. clustering that would be relevant.

- We thank the reviewer for this comment, and we would like to stress out that we completely share their view: network structure is a very important feature to examine given its role in linguistic theories and computational modelling. Since community size and network connectivity are typically confounded in the real world, we agree that any claim made about the unique role of community size in cross-linguistic work should be treated with caution. As such, the goal of the current paper was to make a first step in teasing apart these confounding features by focusing on one of them, namely, community size, and experimentally examine its role when all other features are controlled for (page 2). Accordingly, we chose to manipulate only the number of individuals in the group, while keeping the network structure constant (i.e., fully-connected) across conditions. While differences in network structure could potentially elicit even stronger pressures or modulate the observed effect of group size, examining these additional effects is beyond the scope of the current manuscript. We believe that each of the social features that were argued to play a role in linguistic diversity (e.g., community size, network structure, and the proportion of adult second language learners) requires a separate experimental validation using careful and controlled manipulations, while keeping all other things equal. The results of the current paper show that community size alone plays an important role, and should be considered an important feature as well. Notably, our findings do not preclude the importance of additional factors such as network structure. We now clarify this point in the revised discussion (page 13).
- We share the reviewer's interest in examining network structure effects on language structure. Therefore, we are in the process of collecting data for a new set of experiments that test different types of networks using the same paradigm reported in the current manuscript. In fact, the reviewer's suggested experiment is quite close to the study we are currently running in the lab. We will now outline the new study, and also provide some preliminary results based on the data collected so far. Surprisingly, these results suggest that network structure did not affect the emergence of linguistic structure in the current design.

- In the new study, we are contrasting groups' performances in three different network structure conditions (see Figure below): fully connected networks (in red), small world networks (in blue), and scale-free networks (in green). All groups are comprised of eight participants.



- Our goal is to compare the performance of these different networks using the same measure reported in the current manuscript: linguistic structure, convergence, stability and communicative success. We haven't completed data collection, but following the reviewer's comment, we ran these analyses on the partial data set. Below we provide preliminary results based on the partial data set. The results show a null effect of network structure on linguistic structure (see figure below): while all networks show a significant increase in linguistic structure over time, they do not differ in the speed or trajectory of increase, and all networks show the same degree of high structure by the end of the experiment. We also find similar patterns of results for convergence, stability and communicative success: across all measures, there was no significant difference between the three conditions.



- According to these preliminary findings, network structure does not seem to affect the formation of artificial languages in the lab: All networks reached similar levels of linguistic structure, convergence, stability and accuracy. This result is surprising, but nevertheless in line with the agent-based model reported in Spike (2017), which concluded that network structure plays a relatively small role in the development and maintenance of linguistic complexity when networks have small-world properties.
- Nevertheless, it could that network structure is an important feature which affects language structure, but that we did not capture that for various reasons. One possibility is that our models were not based on sufficient data, and were lacking sufficient power to significantly detect an effect. Another possibility is that our eight-person networks are simply too small to create meaningful differences between network types, but that bigger networks (e.g., of 20 people) will show an effect. That is, it is possible network structure interacts with group size in complex ways (as suggested by Lou-Magnuson & Onnis, 2018). Disentangling these interactions in the lab would require additional conditions, and could be addressed in future work.

Finally, as the paper address compositionality in human language, as opposed to animal communication, I would like to see the notion of compositionality treated in greater detail. Specifically, I would like to see mention of the difference between syntactic and morphological composition and the relation their experiment has for these two domains. For example, is the language game played by the participants capable of capturing this distinction? If not, how might it be altered to address the issue?

- While the difference between syntactic and morphological compositionality is an interesting one, we are not able to capture this distinction in our current experimental design. In our current paradigm, there is no meaningful distinction between sentence-level compositionality and word-level compositionality: complex descriptions in the artificial languages could be interpreted as single words with different affixes, or alternatively as different words combined to a form a sentence (e.g., with a noun describing shape and a verb describing motion). While some participants used a hyphen to separate the sub-strings, we cannot be sure how these sub-strings were perceived and represented. We have added a footnote to the main text to clarify this issue (page 10).
- We speculate that it is indeed possible to improve the paradigm so it is able to capture the distinction between syntactic and morphological compositionality. In the current design, we allowed participants to separate characters only by using the hyphen (which

is indeed ambiguous in this sense), yet future work could also allow participants to use a space (which is a clear word segmentation marker in the Latin Alphabet). Another possible way to make this distinction would be to explicitly ask participants at the end of the experiment how they represented the labels in the language.