Dear Professor Khattab,

I am very grateful to you and the two reviewers for your constructive and thoughtful comments on this manuscript. I found the criticism to be insightful and I have done my best to address each comment thoroughly. Below you will find a copy of the reviewer comments in green, followed by my responses in black, with newly added text *italicized*, where relevant. The relevant sections are highlighted in yellow?? in the main text.

Please note that my Laing (2023) paper has now been published, and so I have updated the reference throughout to reflect this.

Sincerely,

Catherine Laing

AE comments

Dear Dr Laing,

Thank you for submitting your manuscript entitled “Systematicity over the course of early development: an analysis of phonological networks” to Language and Speech. The reviewers note that the study is interesting and highly relevant to the special issue due to the way it tests Vihman’s central proposals on early phonological development. It is also well-written and organised. Each of them has recommendations for revisions that will make the network analysis you have carried out clearer to the reader and that will illustrate and clarify the data and analyses you incorporated into the study. In particular, reviewer 1 ask various questions regarding whether the children’s actual production or target data was used for the analyses in various places and asks for illustrations of networks with real examples. Reviewers 1 and 2 ask that you specify how systematicity is conceptualised in your study and provide a rationale for including children who use two languages. Please refer to the other detailed questions on the analyses and interpretation as well.  
  
I encourage you to resubmit the manuscript taking the reviewers’ comments into consideration.

Thank you for your kind remarks about the paper being interesting, relevant and well-written; I’m glad for the opportunity to strengthen the manuscript further.

Reviewer 1 comments  
  
This study examines systematicity in the vocabulary and early word productions of 9 children (English- and French-speaking children) followed from word production through to 2;6. The author conducts network analyses (network graphs) and uses measures such as mean path lengths and clustering coefficients to determine whether early lexicons consist of closely connected clusters of similar-sounding forms. The author finds general support for systematicity in children’s developing lexicons. Children’s early productions were closer in phonological distance and formed denser clusters than simulated random networks and networks of the target phonological form. Children’s actual productions had higher connectivity than adult target productions.  
  
This paper contributes nicely to the Vihman special issue because it uses sophisticated analysis procedures to test some of Vihman’s central proposals on early phonological development. It joins previous studies by the same author who has used network analyses to examine children’s early vocabularies and word productions. I am not an expert in network analyses; however, from what I can judge, the methodology and statistical analyses of the study seem rigorous and appropriate. The paper is well organized and the writing style is clear. My concern lies in the readability of the study for readers who are not experts in network analyses. It would be helpful for the author to provide more explanations and graphics of what networks and clusters actually look like. It would be useful to see examples of child productions. A summary of major and minor comments is presented below:

Thank you for this positive feedback – I’m glad R1 found the paper rigorous and clear. I will be very glad to further clarify aspects of the paper that may be unclear to a general readership.   
  
Comments  
1.      Since systematicity is an important concept in this paper, it would be worth defining it in terms of how it is used here. Although I appreciate Marilyn Vihman’s work, I don’t think she is the only person who has studied systematicity.

To address this comment, I have added the following text on p.XX:

*“Systematicity is apparent across many linguistic domains, from phonology to syntax, and is noted as being pervasive in linguistic structure [@dingemanse\_arbitrariness\_2015; @nolle\_emergence\_2018]. It is suggested that systematicity may have evolved through transmission over generations, as linguistic structure becomes increasingly ordered through use [@kirby\_cumulative\_2008]. This may be a "design feature" of language that makes it easier to acquire and transmit, as has been shown for adults [@kirby\_cumulative\_2008; @raviv\_what\_2021], children [@raviv\_systematicity\_2018], and in computational research [@monaghan\_design\_2011]. Systematicity may be thus be important in supporting the cognitive processes required to acquire a first language.”*   
  
2.      This study is often compared to that of Laing (under review) and seems somewhat in its shadow. It would be nice for the authors to say more about the unique contribution of this article.

Thank you for this comment. I have emphasised the contribution of the paper more clearly in the literature review on p.XX:

*“This paper contributes a novel approach to the study of developmental vocabulary networks by analysing network \*graphs\*, rather than network \*growth models\* to study phonological networks in the developing vocabulary…and analyse the possibility of a new word being added to the network at the next time-point, rather than the static properties of a network at a given time-point. Network graphs, on the other hand, allow us to understand more about the properties of the network at a given time-point: how ordered/random the network is, how dense its clusters are, and how closely connected words are to one another across the network.”*

Add unique contribution to discussion as well  
  
3.      The difference between a network growth model and a network graph is not so apparent to me and could receive more explanation.

I have also clarified this in the same text as copied above, reiterated here:

*“Network growth models analyse connectivity (are two words similar, yes or no? if yes then they are connected in the network), rather than phonological distance (\*how\* similar are two words in the network?), and analyse the possibility of a new word being added to the network at the next time-point, rather than the static properties of a network at a given time-point. Network graphs, on the other hand, allow us to understand more about the properties of the network at a given time-point: how ordered/random the network is, how dense its clusters are, and how closely connected words are to one another across the network.”*

*Clarify in methods as well?*  
  
4.      Many times it is not clear to me whether the author is referring to analyses done on the target items (vocabulary) or the child’s actual productions. For example, on the bottom of p. 7, the author writes “To address the second question, network graphs of infant’s actual productions will be compared with those of the target form ….”. Hence I assumed that Research Question 1 deals with the target vocabulary. However, it becomes clearer later that question 1 also deals with children’s actual productions.

Thank you for flagging this up. I have now clarified this in the research questions and in the follow-up paragraph on p.XX:

*“1. How systematic are early word productions (both actual and target), and (how) does this change over time?*

*…*

*To test these questions, network graphs will be generated using the \*igraph()\* package [@R-igraph] in R [@R-base] for both the actual and target data. To address the first question, properties of the graphs will be analysed to determine 1) how closely connected individual words are to one another; 2) how dense the overall distribution of words is in the network; and 3) how/whether this changes over time. Following Vihman's work, and findings presented by Laing [-@laing\_phonological\_2023], it is expected that the early vocabulary will become increasingly systematic over time. This would be reflected in denser clusters of phonologically-similar forms and shorter distance between words. Following previous research [REFS], this should be true for both actual and target data. Simulated networks will be used to compare the real networks against both highly systematic and random networks to determine the extent of systematicity present in the data, and developmental changes over time.”*  
  
5.      As mentioned above, it would be really helpful to see some examples of clustering/networks of both vocabulary items and actual productions. Indeed there are no examples of phonological productions in the entire article. I could imagine that the individual networks or clusters are complicated; however, some graphic examples would be useful. In the same vein, can the author illustrate what mean pathway looks like?

Reviewer X and X made similar comments, and I agree it is helpful to visualize this data! To address this, I have added sample network graphs from one of the American infants (Lily) at 14 months (Figure XX), showing Actual and Target networks. I have also added a corresponding table (Table XX) to show the IPA transcriptions of these words in their Actual and Target forms.  
  
Page-by page comments  
p.5, 2nd paragraph – “Other studies have used similar methods to test different kinds of data, to generate consistent (and some inconsistent) results” This sentence seems very general. Does it add anything?

I have changed this to “*These methods have already been used to analyse vocabulary networks across different kinds of data*”, in order to introduce the review that follow in the paragraph.

p.6 “but uses methods that have high potential to do so” – I am not sure why the author adds this phrase since in the previous sentence the author criticized the methodology of the studies, suggesting they couldn’t address systematicity in development.

I have deleted this phrase from this section.  
  
p.7, section on Research questions – At this stage, I am not sure if I know the difference between network graphs vs. growth algorithms.

I hope my response to comment 3 above clarifies this now.  
  
p.8 – Did using only words featured in the CDI result in the extraction of many items? What percentage?

It did! I have now clarified this in the text on p.XX:

“ADD”  
  
p.9 - In calculating the distance values, are the authors using the phonetic form of the target word, the phonetic form of the child’s production or both? This is not clear to me.  
OK – this is answered on the top of p.10. Perhaps this could have been mentioned earlier to prevent confusion.

I hope this is now clear following the revisions I made to address point (4), but please let me know if this is not the case.  
  
p.10 – “The final dataset includes 3223 word types …” The authors are only including word types from the MCDI, so this number would be an underestimate of the total number of word types or not?

I’ve attempted to address this point more clearly in the section on p.XX mentioned above: “*XX% of words in the dataset were excluded due to not being on the CDI. While filtering the data in this way makes the analysis easier to compare across similar studies that also use CDI vocabulary measures (and provides a more manageable dataset for the analysis, given the computational load of comparing all words with all other words in the data), the loss of so many words from the infants’ vocabularies means results likely won’t capture all relevant aspects of the infants' early production.”*

“On average, infants produced 47 tokens of each word type in a single session”. I don’t think I understand this. If the word type was “dog”, children said “dog” on average 47 times in a session? This seems very high. I have never seen this degree of repetition in my recording sessions.

Thank you for spotting this error! It is now fixed and makes more sense: “ADD”  
  
p.12 – RQ1 – From the description of the research questions on p.3, I assumed that RO1 looked at the target words children produced and not their actual productions. Is that the case?

Again, I hope this is now clarified above.

p.13, “Again there was no effect for Corpus on the data”. No mention was made of corpus in the discussion of mean path length.

This has been amended in the text on p.XX: “*there was no change in systematicity over time (i.e. no effect of Age was observed); there was also no effect for corpus.*”

p.13, later “network properties of the Real(Actual) data …compared to the Real Target data”. So the first is the child’s production forms and the second is the target word’s phonetic forms?

I have now amended this to “*network properties of the Real (Actual) data analysed above were compared to the Real Target data - that is, the phonological distance (taken from the IPA transcriptions) between the child's Actual production and its Target counterpart was analysed.”*

p.15, line 7 (counting from the top) – been -> between, e.g., low distance between Actual and Target forms.

Thank you, I have now corrected this.  
  
p.16 – I have difficulty understanding Figure 5. Why does the x axis show negative values?

Re-run and check this  
  
p.17, 2nd paragraph – is “systemtatic” a word or should it be systematic?

Thank you – that was a typo and has now been corrected.  
  
p.18, end of 1st paragraph – what is the significance of only network size vs. age predicting learning?

Having re-read this section I think the original text was a little misleading, and so I have revised this as follows, which I hope clarifies this more successfully:

*“Looking at the comparison between Actual and Target data, neither mean path length nor clustering coefficient changed with age. However, network size was a predictor of both effects, but in the unexpected direction: higher mean path length and lower clustering coefficient were observed as network size increased. This is contrary to what was found by Laing (2024) and appears to indicate a decrease in systematicity over time. These contrasting findings are likely driven by differences in what was being analysed: the network growth algorithms in Laing’s (2024) paper predict the likelihood of a given word being added to the network in the next month, and, over time, the network was more likely to acquire new words that would connect to the most densely-connected words in the existing network (controlling for network size). Adding the present findings to this picture, the densest clusters at later timepoints were in fact less dense than those at earlier timepoints, with larger phonological distances between connected words. This aligns with what we know about systematicity over the course of early phonological development, as the kinds of words being targeted for production become more variable. This is demonstrated in case study accounts of infants' early words, where we see the establishment of different production patterns, or templates (Vihman, 2019) over time. This is clearly demonstrated in Waterson's (1971) case study of her son's production, for example, where five distinct structures are identified in his data, to which newly-acquired words are systematically adapted. In this example, we see systematicity becoming more prevalent in the data (as a wider range of templates gives rise to more opportunity for word adaptation) but clusters of similar words may be less dense, as adaptation takes place in a number of different - but systematic - ways.*  
  
p.19, bottom of page – I have difficulty following the argumentation. In essence, the author is refuting her previous claim?

Following the comment above, I have now clarified and merged these two paragraphs as they were essentially both saying the same thing.   
  
p.20, 3rd line – care study -> case study  
  
Thank you – this typo has now been corrected  
  
  
Reviewer: 2  
  
Comments to the Author  
The authors contribute an interesting study of systematicity in early words with a focus on two key data sources (Providence: Demuth et al., 2006; Lyon: Demuth & Tremblay, 2008).  By using a network modelling approach, they strive to better understand the extent to which systematicity underlies early phonological development.  The manuscript is generally well-written and organized.  The suggested revisions focus on bringing greater clarity and specificity to the way that they converted the words in these data sources to be incorporated into the study.  This clarification is essential to interpret the findings and consider the implications.  The study included data from English and French but did not justify this choice nor was there a consideration of the implications of these findings in a cross-linguistic context.  Lastly, a few more details would be required to orient the reader to this analysis approach.  Please see the detailed feedback below.

Thank you for taking the time to review the paper, and for these positive comments; I am glad R2 found this paper well-organized and interesting.

page 7: “This work thus cannot address any questions about systematicity in early acquisition (though note that this was not the intention of either paper), but uses methods that have high potential for doing so.”

* In the review of Vihman’s work, the concept of systematicity is introduced. Consider adding in that section to outline what are the conditions needed to study systematicity in early acquisition. Based on the critique of the Fourtassi et al & Siew & Vitevitch, it seems that we would need (a) children’s phonological productions of early words, (b) longitudinal data (is this to understand the development of systematicity? Could systematicity at a single timepoints be analyzed?). Are there other (minimum) requirements? Sufficient size of sample? This information would help frame the study.

Say more also about Kalinowski study

page 7: “This study expands on previous work, and builds on findings from Laing (under review), by analysing network graphs of infants’ early lexicons.”

* While I understand the general concept, I’m still new to network models. For readers like me, could you expand here to make clear how network graphs expands on previous work that used network growth algorithms (from my reading, this seems to be what the paper is building from)?

Same as R1’s Q

page 8: “Following Vihman’s work, and findings presented by Laing (under review),…”

* Since the present paper builds on Laing, have you considered what re-writing may be needed if the paper is published before the review process is completed? Is there a way to associate these two works by pointing to other work by Laing (e.g., conference proceedings, working paper)?

First paper is now published – need to check findings are reported accurately in the text

page 9: “This was drawn from two corpora on PhonBank (Rose & MacWhinney, 2014): Providence (American English - Demuth, Culbertson, & Alter, 2006) and Lyon (French - Demuth & Tremblay, 2008).”

* What is the motivation for including children who speak 2 languages? On one hand this may speak to cross-linguistic generalizability, but on the other may add variability to the data?

Consider incorporating this as a research question.

Increase sample size  
Cross-ling generalizability  
No expectation that there should be different results across languages (may be lang-specific tendencies in early production, but the data captured here doesn’t account for that); we shouldn’t see differences according to past research

Script written initially for English but wanted to figure out how to adapt it to different languages; opens door for other researchers to analyse data across more languages

page 10: “Distance values were established using methods set out in Monaghan et al. (2010), using distinctive features to generate a set of phonetic values for each word that could then be compared with all other words”

* Could you expand on this distinctive feature analysis (include an example, or more explanation) as this calculation was key to the analysis

Include in supplementaries? Or just refer back to Laing?

page 10: “Euclidean distance between the values of each word and each other word in each infant’s global network was then used to determine how close/distant words were from one another.”

* the distance was with regards to the distinctive feature value generated for each word? It would seem that words could have similar distinctive feature value, yet be composed to quite different phonemes?

This is a good point; this risk is mitigated to some extent by the fact that comparisons were done on a segment-by-segment level, so the 14 distinctive features were compared across every pair of segments. This means that sonority was always compared with sonority, voicing with voicing, etc. That being said, and I think speaking more directly to R2’s point, a pair of very similar words could potentially have the same distance from one another as another pair of less similar words. To test this…..

page 10: “Often, infants produced multiple tokens of the same word type in a given month, often with high variability across tokens.”

* what kind of variability? Variability in distinctive feature value, or in the consonant productions themselves?

Add to clarify - perhaps include examples?

page 10: “Because it was not possible to generate networks with all tokens included…”

* please specify what “tokens” are (children’s words? Consonants?)

This has now been changed to “all word tokens included”.

page 10: “a mean value for each distinctive feature was established across tokens, meaning that each word’s distinctive feature value represents the variability of the infant’s production of a given word.”

* Please provide more explanation. From this description, I am not sure what was calculated and makes the interpretation of the results difficult, and replication of this approach very limited.

Add an example

* page 11: Re. Distance score of .25
* How was the distance score chosen? Are there guidelines for making this decision?

Refer to Laing study

* page 11: “The final dataset includes 3223 word types in total... “
* Consider adding a table that provides an overview of the total number of tokens over time per child (e.g., per month as that was how the network graphs were completed) as there was variability within and across children in these datasets that would be interesting for the reader to be aware of.

Replicate from Laing study?

\*page 11: Re. mean path length indexes the average phonological distance

* consider changing “average phonological distance” to something that more closely reflects the measure (e.g., “average distance of all consonants in word” “average distance in distinctive features” ?)

The measure of phonological distance is taken from counting Euclidean distance between segment pairs across a word, and so it is a bit too reductive to go with average distance between xyz.

Maybe add something to the paper: “Euclidean distance between the values of each word and each other word in each infant’s global network was then used to determine how close/distant words were from one another (hereafter phonological distance).

…

Two key variables will be explored through an analysis of network graphs: *mean path length* and *average clustering coefficient*. These have been calculated from a measure I refer to here as phonological distance, which is derived from the Euclidean distance…”

* page 15: “To address the second rese"arch question, the phonological distance...”

I don’t understand this comment

page 16

First, GAMMs were used to examine connectivity of the infants’ Actual and Target networks and how these changed over time.

* The description of GAMM and model (last 2 paragraphs of page 15 first 2 paragraphs of page 16) would better fit in the methods section

Easy to address

page 18: “Overall, infants’ early productions were closer in phonological distance (mean path length) and formed denser clusters of similar forms within the networks (average clustering coefficient) than simulated random networks and networks of the target phonological forms, though these were less systematic than prototypical highly systematic “small world” simulated networks.”

* While this research is very interesting and analytic approach quite novel, I would like to see a more careful interpretation of the results that stay closer to the measures used. Please reconsider “phonological distance” to reflect better the data since there were a number of ways that the data was transformed: (mean distinctive features, focus on consonants only, proxy of selection/adaptation).

I think an exploratory overview of words that are similar in phon dist might support this; return to this once I’ve done that.

page 20: “The consistency in evidence for systematicity in the data, in particular in the Actual forms, lends strong support towards the argument for a reliance on a “phonic core of remembered lexical items and articulations” (Ferguson & Farwell, 1975, p. 112) in development, which are systematically drawn upon to tackle the challenges of remembering and producing early words.”

* I need a better understanding of how the the phonological forms of the words were represented in the model before considering the extent to which this data supports this argument (and the presence of systematicity via actual forms).

As above

page 20: “On the other hand, a key difference across the results is the contradictory findings regarding change over time. However, to address this we might consider what it would mean for these two variables to change in the opposite direction over time (i.e. indicating an increase in systematicity).”

* These hypotheses could be more clear from the outset of the manuscript. That these two variables change in opposite directions indicate increase in systematicity? If this is the case, this also could be made more clear in the introduction when the Laing study is discussed.

Need to read the paper and review this comment

page 21: “Indeed, this outcome aligns well with care study accounts of infants’ early words, where we see the establishment of different production patterns, or templates (Vihman, 2019) over time.”

Is “care” an extra word in this sentence? If not, I don’t quite understand what it means in this context.

Apologies, this was a typo and has now been corrected to “case study”.

page 21: “This is clearly demonstrated in Waterson’s (1971) case study of her son’s production, for example, where five distinct structures are identified in his data, to which newly-acquired words are systematically adapted.”

* what about the age range under study in Waterson (seems to be mainly at 1;6); could the lack of change with age be related to the age of children in the present study? Or less driven by age but more by vocabulary size (which varies quite a bit within this age range?)

Review this

Next steps:

* read the paper
* revisit the analytical comments
* review the number of tokens per infant/session
* check refs to JExP paper are still accurate
* review the data to determine the extent to which the phon dist measure is representative