Peer Review Comments

Article: Lockwood, G, Hagoort, P and Dingemanse, M 2016 How Iconicity Helps People Learn New Words: Neural Correlates and Individual Differences in Sound-Symbolic Bootstrapping. *Collabra*, 2(1): 7, pp. 1–15, DOI: http://dx.doi.org/10.1525/collabra.42

Article type: Original research report

Subject: Psychology

Editor: Rolf Zwaan and Max Coltheart

Article submitted: 06 April 2016

Editor decision: 26 April 2016, 25 May 2016

Revision submitted: 12 May 2016, 17 May 2016

Article accepted: 25 May 2016

Article published: 06 July 2016

Responses for Version 1

Reviewer: This reviewer has chosen to remain anonymous
Affiliation: --
Competing Interests Statement: No
Review Completed: 20 April 2016
Recommendation: Revisions Required

1) General comments and summary of recommendation:
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?

This is an interesting paper, investigating the fascinating topic of ideophones in language learning. I cannot comment on the EEG part of the study, but I can comment on the behavioural experiments. The authors trained Dutch participants on meanings of Japanese ideophones in two conditions: (1) with their matching translation, or (2) with a non-matching translation. Participants scored more accurately on the posttest for ideophones in the matching than the non-matching condition. Participants also showed relatively high accuracy in choosing the correct meaning of ideophones in a second task even if they had previously learned a non-matching meaning. The authors investigate how individual differences modulate learning in further analyses taking ERP data into account.

I had two main issues concerning the methodology:

1) It wasn't explicitly stated whether the design was fully counter-balanced across items. If it wasn't, there is a chance that results are not generalizable to all ideophones (see details below).

2) I wasn't convinced that it had been established that the majority of participants were above chance when choosing a meaning for ideophones in the sound-symbol sensitivity check. This is particularly important for items previously learned in the non-matching condition. Obviously, only if performance is significantly above chance can it be taken as evidence for the iconicity for ideophones.

I will comment on the paper by section below and also provide more details regarding these two points. I will suggest several changes for a revision of this paper.

Introduction & Background:

-I thought that the Introduction would benefit from the addition of relevant theoretical frameworks. What cognitive/linguistic/brain processing theories does this work speak to?

-P.3: I think it might be good to spell out explicitly what kind insights into cognition and/or brain processes you were hoping to gather by carrying out imaging work in this area, and why the type of imaging technique you chose is best placed to do this.

Stimuli:

-I might have missed this, but if I understand correctly, the design wasn't fully counterbalanced, such that the 19 items in the 'real' (matching) condition for some participants were in the 'opposite' (non-matching) condition for others? I think that what was done instead is that all participants learned the same 19 items in the matching condition. And they learned the remaining 19 items in the non-matching condition. While words across conditions were matched (please provide a table with relevant matching data), I worry that by not using the same words in both conditions, we cannot be sure that it wasn't something about the 19 items in the opposite condition that made them harder to learn, rather than the fact that they were being taught with their opposite meaning. Given the model comparison showed a random effect by ideophone (p. 8) this possibly confirms that there are differences between items despite the fact that they were selected so carefully. If the design wasn't fully counterbalanced, this needs to be acknowledged and where statistical procedures can partly eliminate any bias this should also be made explicit.

Procedure:

-What were the instructions to participants?

-If I understood correctly, in the testing phase, participants saw each word twice: once with the previously learned Dutch meaning and once with another meaning. In this sentence on page 8 "participants correctly remembered the real word pairing 86.7% of the time", does "correct" mean the combination of responding YES to the learned meaning and NO to the other meaning?

Participants:

-I had to read through these sections several times to work out how many participants ended up in the final analyses. Attrition was different across the different aspects of the study (behavioural and ERP). Hence, a definitive sentence on how many participants' complete datasets across the different parts of the study were included in the final analyses would be quite helpful.

-I take it the participants were naïve to the Japanese language? Please mention briefly.

Behavioural Results: Main experiment (learning task):

-Given the authors chose to depict individual data (which I think is a great choice), it is possible to see from the Fig. 4 that there is not only variability in spread between the two conditions, but also that not all participants showed the advantage for the "real (matching)" over the "non-matching" condition. The authors describe individual difference data in their work which is great. Can this be extended to the following: How many individuals show the main pattern (advantage for learning ideophones with correct translation)? Wilcoxon two sample test should be fine to use here (accuracy and RT).

Sound-symbolism sensitivity check:

-Section 3.1.2 seems to move back and forth between results by participants and by items. This made it a bit hard to follow what was going on. Maybe the authors could state guessing accuracy (and whether this was above chance) first by participants, then by items (or whichever order works best)?

- You currently state that one participant guessed at 50% (presumably you mention this because it's an at chance performance). However, if I have this correct, then an above chance performance is more than 24 correct out of 38 items on a two alternative choice (using a one-tailed test) which would be 63% correct items or more. Could you check the above chance level performance and report how many individual participants scored above chance?

-Initially, participants had learned 86.7% of the words in the correct condition. In the sound-symbolism check they dropped down to 77.3%. It is unclear to me why they would perform worse here. If learning effects and the natural sense sound-symbolism makes to people adds up, participants should

have got even better. But maybe these effects don't add up. Maybe this condition is too difficult to interpret – many things could be going on (maybe participants second guessed themselves because they had just been informed that some [but not which] words they had learned actually had a different meaning). What did the authors hypothesise the results would be and why? In any case, it seemed to me that the interesting comparison here was the non-matching condition – learning task vs. forced choice sound-symbolism sensitivity check: Participants' accuracy on words from the non-matching condition was 71.3% in the learning condition (that is, having learned the incorrect pairings) to 68.6% correct responses (that is, the actual meaning) in the forced choice task. I thought that it is interesting that participants' sensitivity to sound-symbolism allows them to achieve a level of accuracy that matches their learning ability. Again, it is important to establish that accuracy levels in the forced-choice task are above chance level.

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, if relevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).

I found the figures clear. It is great to be able to view individual data (e.g., Figure 4)

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.

Ethical approval statement provided.

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.

I found this paper well written and generally easy to follow. Below are some specific comments on aspects I found somewhat harder to understand.

1) I suggest explaining the terms iconicity, sound-symbolism and ideophones (including giving examples) to help make this work more accessible.

2) Make sure to introduce acronyms in brackets before you start using them (e.g., p.4).

3) P. 3 (last paragraph): Expression "speed of reaction" suggest to replace with "response time"

4) P.3 (first paragraph): Description of Lockwood et al study. I had some difficulty following the description of the second learning study with 'regular' adjectives. First, what do you mean by 'regular' adjectives? Non-ideophones? Please clarify

5) P.10: change "reëvaluate" to "re-evaluate" (2 occurrences)

Reviewer: Nic Badcock
Affiliation: Macquarie University, Australia
Competing Interests Statement: I have no competing interests to declare.
Review Completed: 18 April 2016
Recommendation: Revisions Required

1) General comments and summary of recommendation:
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?

- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?

Overview:

The works compares the learning of Dutch and Japanese word pairs in two conditions, manipulating the matching of the meanings: in one condition the meanings were matched, in the other condition the meanings were opposite. Critically, the Japanese words are ideophones that offer some sensory imagery in their sounds (guessing at ideophonic word meaning is typically above chance). The work was conducted in three phases, learning, testing, and symbol-sound checking. Behaviourally the (Dutch) participants were more accurate at learning words for correctly matched meanings and were able to guess the true word meanings better than chance for all stimuli, irrespective of learning condition – replicating previous research.

The novel aspect of the works was the including of EEG to assess event-related potentials (ERPs) to the stimuli. The was a relationship between P3 and later ERP waveforms to the ideophonic stimuli in the checking phase, which (through careful analysis) was revealed to be related to learning phase accuracy. The pattern indicated that greater sensitivity to the sound-symbolism was related to poorer accuracy learning word pairs with opposite word meanings.

Dear Mr Lockwoord and colleagues,

Overall this is a very nicely written and engaging piece of work – I really enjoyed reading it. To address the Collabra guidelines, I think the methodologies are appropriate and I didn't pickup any major weakness to be addressed. The work is logically structured, succinct, and the argument flows coherently – though I would like to see a little more rationale for the introduction which I'll detail below. I also found the references to be adequate and appropriate, bearing in mind that I have limited knowledge of the language side of this project.

1. Is all statistical analysis sound?

This is really a minor comment. One thing I wondered about was testing for a group interaction related to ERP ANOVAs run separately for the two groups (page 18). It would make for a stronger case if the interaction were statistically reliable but I appreciate that this may not come out. I don't see that this would detract from the work. Considering the exploratory nature of the EEG component of this work, whether or not the interaction is significant, the results do beg for more work to be done

- Consider updating the ANOVA to include a group factor

2. Does the conclusion (if present) reflect the argument, is it supported by data/facts?

This is again minor. For the most part, I found that the discussion and conclusions fit well with the data however, the strength of the relationship feels to be overstated (page 19).

- Consider rewording to better reflect the magnitude of the associations

3. Is the article logically structured, succinct, and does the argument flow coherently?

As with the previous comments, I only had a minor concern with the flow of the argument and this related to the introduction/rationale for the EEG side of things. On page 4 you mention that you had no hypotheses regarding the ERP components, and similarly, you talk about relationship between the ERP effect size and behavioural measures without context. EEG is relative intensive thing to include if it's not important, and I would like to see a little more explanation of why EEG it was considered necessary at the outset.

- Please include more of an introduction as to why EEG was included

- In relation to this it would also be useful to have some background/rationale regarding the sites chosen for analysis.

Past literature is referenced regarding the P2 as a time window of interest and as it doesn't jump out of the results it isn't featured beyond the methods. I think it's important to return to this in the discussion and it's currently one of the main motivations for the work.

- Please provide some discussion of the relevance (or lack of) of the P2

Finally, following the break-down of the P3 effects in relation to individual differences (page 18), I found myself expecting a similar analysis for the later positivity but this didn't occur. Again, in fitting with the previous literature that provided the motivation to include EEG, I think it's important to follow through with this analysis.

- Consider including an analysis of the individual differences with respect to the late positivity

4. A limitations section

Though not related to any of the Collabra guidelines, I think this might be some merit to including a limitations section for this work, particularly to highlight the exploratory nature of much of the analysis. I don't see this as a negative but I do think it's useful for interpretation of the work.

- Consider including a limitations section

I've included a series of specific comments below, many of which highlight particular points that motivated the above comments. A number of these relate to more complete reporting of statistics, including non-significant effects, and effect sizes throughout, as well as clarifying some of the presentation and processing procedures.

Best wishes,

Nic Badcock

Specific comments:

5. Page 3 "…Lockwood & Tuomainen (2015) found that ideophones elicit a greater P2 component and elicit a larger late positive complex compared to arbitrary words."

- I noticed that LPC occurs later in the text and would be nicely defined in this sentence. Having said this (upon my first read), after finishing the paper, it didn't feature heavily so may not be a necessary abbreviation – please consider whether it's needed

6. Page 4 "…we made no formal hypotheses." Regarding the EEG measures

This seems atheoretical – feels like there should be some reason to wire-up the participants. This is just a comment here – no need to respond but I'll include some more thoughts in the major comments above.

7. Page 4 "We also investigated the relation between the ERP effect size, the memory/learning performance of the task, and behavioural measures of sensitivity to sound symbolism per participant."

- It would be useful to provide some rationale for these comparisons

8. Figure 1, 2, and 3 timing

- Suggest including the timing information as part of all figures – just allows readers to extract the experimental details more easily

9. Page 5 "The initial Dutch word was presented for 1000ms with 100ms of jitter, …"

- Please clarify what the jitter does in this instance – that is, it is 950 to 1050 ms, 1000 to 1100 ms, or something else? It'll be important to have this for replication

10. Page 6 "Between trials, a blank screen was presented, followed by a fixation cross to announce the beginning of the next trial."

- Please include the timing of the blank screen and the trial initial fixation as well

11. Page 7: "… recruited from the MPI participant database."

• Please define and provide some background on the MPI database – e.g., general population versus students

12. Page 7 "We first used non-parametric cluster-based permutation tests to investigate the entire epoch, and then ran ANOVAs on individual time windows of interest."

• Please include more explanation of why and how these steps were done. For example, regarding the time window analyses – where these based on peak amplitudes or window average amplidue?

13. Page 8 "This is shown in Figure 4 below, presented in this way rather than as a histogram with error bars in order to better represent the spread of data"

This is a great way to present the data! Really good to be able to see it all. Just one suggestion to add to this.

• Consider including means and 95% confidence intervals to clearly map onto the statistics presented in text

• And/or consider presenting the confidence intervals in text accompanying the means.

14. Page 8 "…to ideophones in the real condition (mean RT = 958ms) than the opposite condition (mean RT=1262ms) ($t = 6.12$, $p < 0.001$)"

• Please add some indication of variance to the mean RTs as well as a measure of effect size for the t-tests (including the next sentence, and for the sound sensitivity check in section 3.1.2)

15. Page 8 "However, there was no correlation between the number of correct responses per participant… "

• Please include the statistics for the non-significant relationship between accuracy and reaction time. Might be useful for a meta-analysis at some point in the future

16. Page 10 "… at 77.3%, and they guessed the real meanings of the words they had previously learned in the opposite condition at 68.6% …"

• Please include a measure of variance for these values

17. Page 10 "However, they may have found it harder to reverse this learning than they did to reëvaluate the ideophones they had learned in the real condition ($t = 1.9665$, $p=0.057$)."

• Please be explicit about what the inferential statistics refer to here

• and include effect size

• Also note the umlaut typo for re-evaluate (also in the next sentence)

18. Page 12 "but there was no effect when looking at the first learning round, the second learning round, or both together."

• Please include the inferential statistics for these comparisons - might be best in a table so that it doesn't disrupt the flow of the text (very nice as it is)

19. Page 14 "… eliciting greater a P3 ($F=16.99$, $df=1,28$, $p=0.0003$) and late positive complex ($F=8.96$, $df=1,28$, $p=0.0057$)."

• Please include effect size for these inferential stats

• Also for those in tables 2 and 3

20. Page 14 "…and statistical analyses revealed similar results to the analyses of all trials, but all effects were weaker due to having fewer trials."

• Consider including these statistics at some point – even if it's in supplementary materials. Again, might be relevant for a meta-analysis at some point so nice to make sure it's available at this stage of the process

- Similarly for the handedness check

21. Page 16 "These measures were significantly correlated (r=0.42, p=0.0236), meaning that participants who are better at guessing the meanings of ideophones show a greater P3 effect."

- Consider including a scatter plot of this relationship

22. Page 18 "…for the bottom 14 (F=4.13, p=0.063) while remaining consistent for the top 15 (F=14.30, p=0.0020)."

- Be great to have effect size here

- It would also be nice to see the interaction for group here – so running a bigger ANOVA to check the statistical reliability of this difference

23. Page 18, paragraph 2 – P3 by group description

This is really interesting but I found that I needed to read it a few times to get my head around it. I wondered whether it might be easier if you abbreviated the groups in some way. I don't have a suggestion that I'm particularly happy with at this stage but something to clarify this would be helpful.

- Consider rephrasing this paragraph in someway for clarity

24. Page 18, Late positivity

In reading this section, I anticipated seeing a similar description/comparison for the later component of the ERP. If there is some theoretical reason for separating the two time points, we'd expect to see some differential prediction.

- Consider including a similar section for the late ERP component

25. Page 19 "close relationship between the ERP results and the two behavioural measures of sound symbolism: performance in the sound-symbolic sensitivity check and differences between test scores across conditions. We found that the magnitude of the ERP effects predicted performance in the behavioural tasks and thus serves as a reliable index of sound-symbolic sensitivity."

I think the 'close relationship' and 'reliable index' is overstating the strength of the relationships noted in the data.

- Please consider rewording to better represent the strength of the relationships.

26. Page 20 "This confirms that sound symbolism boosts word learning in adults learning words in a new language as well as in infants and children."

I don't think you can confirm anything about infants and children here

- Please rephrase to clarify the current results from existing evidence

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, if relevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).

I thought the figures were very nicely presented. I could not see in the accompanying files or notes in the text that the data was available, though I did note that the R script were included.

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.

Humans were included and the typical ethical approval information was included in the manuscript

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.

Yes

Reviewer: Margriet Groen
Affiliation: Radboud University Nijmegen, Netherlands
Competing Interests Statement: I have no competing interests to declare.
Review Completed: 21 April 2016
Recommendation: Revisions Required

1) General comments and summary of recommendation:
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?

In their manuscript "How iconicity helps people learn new words: neural correlates and individual differences in sound-symbolic bootstrapping", Lockwood, Hagoort and Dingemanse present a word-learning study using behavioural and event-related potential (ERP) measurements in Dutch participants being exposed to Japanese ideophones (i.e., words that depict sensory imagery). It aims to confirm and extent previous work by 1) replicating behavioural results by Lockwood, Dingemanse & Hagoort (2016) in a new sample, and 2) investigating electrophysiological responses to sound-symbolism.

The study is well conducted and mostly reported in a clear manner. Methodologies, statistics, and references are all sound and the article is logically structured.

I recommend the paper for publication in Collabra after some revisions to the manuscript to increase clarity.

In addition to some minor points summarized in the attached files (both as comments in the pdf readable in Skim, and a summary of the comments in a txt-file), something I miss in the paper is a clear rationale as to what the use of EEG/ERP adds to the behavioural measures. Although I do believe the ERP results are informative, this could be more clearly formulated in the introduction, and expanded upon in the discussion.

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, if relevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).

Tables:

The information in the first two rows of Table 3 is the same as the information in Table 2, making Table 2 redundant.

Figures:

I would prefer the authors to combine several figures into one figure with separate panels and include more informative figure legends to render them less dependent on the text.

For instance, figures 1-3 could be integrated into a single figure with three panels.

Similarly, Figures 4-7 could be integrated into one figure with 4 panels, and Figures 8 and 9 into one figure with two panels. The same holds for figures 12 and 13.

Figure 5: To me it would be easier to read if accuracy were on the y-axis and reaction time on the x-axis. This would also be more consistent with all the other figures reporting accuracy data.

I think figures 10 and 11 could be added as supplemental. They aren't really necessary to understand the results.

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.

Informed consent is declared.

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.

The text is well written and the quality of English is very high.

Editor Decision for Version 1

Editor: Max Coltheart
Affiliation: Macquaire University, Australia
Editor decision: Revisions required
Decision date: 22 April 2016

As one of the reviewers for the submission, "How iconicity helps people learn new words: neural correlates and individual differences in sound-symbolic bootstrapping.," to Collabra, I am sending you the reviews and editorial decision sent to the author of this piece. Thank you again for your important contribution to this process.

Max Coltheart

Macquarie University

max.coltheart@mq.edu.au

Dear Mr Lockwood,

Thank you for submitting your interesting paper on sound-symbol iconicity. It has been reviewed by three reviewers, whose recommendations were identical: the paper is publishable after revisions. The reviews are appended below. Each reviewer has listed desired revisions. There are quite a few of these but I can't see any that would be at all difficult for you to handle, and I do think that attending to them will increase the clarity and readability of the paper. I don't expect to need any further reviewing of the paper.

When you submit the revision, may I ask you also to send me a covering letter listing the requests for revision that the reviewers have made and how you have responded to each or, where you have not responded, why you thought no response was needed.

Best wishes,

Max Coltheart

Author's Response to Review Comments for Version 1

Author: Gwilym Lockwood
Affiliation: Max Planck Institute for Psycholinguistics, Netherlands
Revision submitted: 12 May 2016

Hi,

I've just uploaded the revised version of my manuscript on the revisions page of the submissions section. I've uploaded the following files:

Revised ms as pdf

Revised ms as docx

Response to editor detailing all changes as docx

All new figures in order (although it's hard to tell, as Collabra changes the upload name)

All supplementary materials, including stimuli, sound files, participant data, and analysis scripts, can be seen on the OSF project page at https://osf.io/ema3t/

Please let me know if there's anything else you'd like!

Cheers,

Gwilym

Attached document:

Dear Max,

Thanks for the incredibly quick turnaround and for the useful suggestions for improvement. I've added and/or amended the various sections that you highlight, and I'd also like to explain why certain statistical procedures weren't performed (or reported).

There were three main issues that the reviewers brought up, which were:

1. More rationalisation and justification for the use of ERPs and discussion of the time windows

2. The counterbalancing of the stimuli

3. The addition of a limitations section

The issue of rationalising and justifying ERPs is a good point. On re-reading the introduction, I myself found that section less than ideal; it read like we were unsure of whether to outline previous work on sound-symbolism in general before introducing my work or not. This meant that the discussion of previous ERP sound symbolism research got lost somewhat. We have tidied the flow of the section up to make that a lot clearer. There isn't that much ERP work on sound-symbolism out there yet, and of the studies that have been done, they've found various different time windows of interest. The main point of using ERPs at all is to establish what kind of mechanism sound symbolism is based on; there are arguments about whether it's an early, low-level sensory effect, or something that happens at the linguistic/meta-linguistic level. We have added or rephrased a few parts here, but the main section is as follows:

We used ERPs to look at the time course of the neural effect; if an early effect was present, as in Kovic et al.'s (2010) study with pseudowords, this would suggest that the effect is based on differences in the processing of the sensory properties of the stimuli, whereas if the effect was much later, it would suggest a more linguistic mechanism. It is possible that there are both sensory and linguistic effects, as suggested in ERP experiments by Lockwood and Tuomainen (2015) and Sučević et al. (2015).

We have also clarified the hypotheses, as the original manuscript focused too much on my earlier ERP sound symbolism work while also stressing that the strength of the non-parametric cluster-based permutation test is that it highlights windows of interest in a theory blind way. We rephrased the hypotheses

section as follows:

We hypothesised that we would behaviourally replicate Lockwood et al. (2016), namely that participants would learn the ideophones in the real condition better than in the opposite condition and that participants would still be sensitive to the meanings of ideophones in the two-alternative forced choice task afterwards despite the learning rounds. We also predicted that there would be a correlation between the reaction time and accuracy of judgement of ideophones, in that the more accurately guessed ideophones would also be more quickly guessed. As for the ERP results, since the few sound symbolism ERP studies so far have found different components, we used a non-parametric cluster-based permutation test to investigate the data before analysing particular windows. Finally, we investigated individual differences in the data by looking at the relation between the ERP effect size, the memory/learning performance of the task, and behavioural measures of sensitivity to sound symbolism per participant. We did this in order to see whether the effect was more related to participants' sensitivity to sound symbolism or more related to participants' general task performance.

The issue of counterbalancing the stimuli is definitely something that needed to be addressed. The stimuli were not counterbalanced across participants - each participant learned the same word pairs, meaning that the ideophones which were learned in the correct condition were in the correct condition for everybody, and likewise for the opposite condition. We have added a sentence clarifying this, and outlining it in the limitations section, and also discussing why we did this. Behaviourally, this is almost an exact repeat of the procedure in Lockwood et al. (2016), where the stimuli weren't counterbalanced either. However, in Lockwood et al. (2016), we outlined a pre-test which did use counterbalanced stimuli and found the same effect for both groups. In any case, this is now mentioned in the methods section and the limitations paragraph in the discussion.

The limitations section is a good idea, and we have put it as the final paragraph in the discussion. This covers the stimuli counterbalancing limitations and the exploratory nature of the individual differences measures. It reads as follows:

One limitation of the current study is that the stimuli were not counterbalanced across participants. However, we found in pre-tests with the same counterbalanced stimuli in Lockwood et al. (2016) that the behavioural learning effect was consistent for both groups. Another caveat is that the individual difference data is exploratory and should not be taken as conclusive.

There were also several individual minor points which the reviewers. Our response to 90-95% of the minor points was "yes, good point, let's do that" and changed the manuscript pretty much exactly as requested. I will outline the individual points which we did not address exactly or which require a little more explanation reviewer by reviewer below.

We have also made two extra additions which we realised we missed earlier. Firstly, on page 6 we have added a sentence that describes the consent form with regards to data sharing. We had separate sections for general experimental consent and data sharing consent (i.e. people could participate but refuse to let us share their individual data), and the default was that individual data would not be shared unless otherwise told. All participants were fine with data sharing. The added sentence is:

Participants were told that data sharing was optional, and all participants explicitly opted in to consenting to their data being shared.

Secondly, on page 9 we have added a graph comparing the individual behavioural data from this experiment with the Lockwood et al. (2016) JEP:LMC paper that it is based on. We feel it improves the paper by explicitly showing how closely the results have been replicated, rather than simply saying so. Everything is fine in terms of copyright - it is a completely different graph from the black and white histograms used in the JEP:LMC paper, so reuse rights should not be a problem since there isn't any actual reuse happening.

Finally, we have also added a Supplementary Materials section with hyperlinks to the relevant pages on OSF.

Thanks again for a wonderfully quick response, and to the reviewers for some really kind and useful reviews. It was especially good to hear that people liked Figure 4 (now Figure 2); histograms are so standard that I thought people would dismiss it and ask for histograms, so it was really nice to read reviewers' compliments on it!

Best wishes,

Gwilym

Reviewer A

"What cognitive/linguistic/brain processing theories does this work speak to?"

Our Frontiers review paper on sound symbolism research (Lockwood & Dingemanse (2015), Iconicity in the lab) outlines the relevance of sound symbolism to cognitive, linguistic, and brain processing theories. In that review, we write that one of the reasons that there are no fleshed-out models of sound symbolism processing is because of the paucity of either highly specific behavioural research or neuroimaging research. We see this as an exploratory piece of the puzzle for developing a fuller theory rather than a research paper directly testing a fuller theory.

"How many individuals show the main pattern (advantage for learning ideophones with correct translation)? Wilcoxon two sample test should be fine to use here."

23 out of 29 participants show the main pattern. Of the six who don't, two are equal in both conditions, and four have minor advantages in the opposite condition (with a mean advantage of 3.95 percentage points higher than in the real condition). I've added this to the analysis. However, I'm not certain that the Wilcoxon test tells us anything that the mixed model doesn't already tell us - namely that there is a strong and consistent behavioural effect. I agree that it is good to have a measure of the number of participants showing this effect, so I've added that.

"Could you check the above chance level performance and report how many individual participants scored above chance"?

Using Reviewer A's measure of above chance meaning guessing 24/38 ideophones correctly or more, 27 out of 29 participants guessed 24 or more correctly. We also re-ran the t-test with 24/38 as the mu measure to compare against, and the result was still significant ($p<0.0001$). The 27 out of 29 figure has been included in the manuscript, while the re-run t-test has been included in the R Markdown file for transparency.

"What did the authors hypothesise the results would be and why?"

(referring to the post-test 2AFC task)

We hypothesised that participants would guess the ideophones' real meanings at above chance regardless of which condition they were learned in, and this is already mentioned on page 3 of the manuscript.

"Change 'reëvaluate' to 're-evaluate'"

This isn't a typo and I prefer using a diaeresis to spell English words with double vowels which require hyphenation (re-evaluate, co-operate, etc.). This used to be a lot more common, and the New Yorker still uses it. We'll leave this one to the Collabra editorial team for an executive style guide decision. The same holds for Reviewer B's comment.

Reviewer B / Nic Badcock (thanks for naming yourself!)

"This is really a minor comment ... consider updating the ANOVA to include a group factor"

I did this, and the interaction wasn't significant. As Nic said that it wouldn't detract from the work, we included it in the R Markdown file, but not the main manuscript.

"Consider including an analysis of the individual differences with respect to the late positivity"

Correlations and some plots are already provided in the R Markdown file.

"Consider including these statistics [page 14, referring to correct-only trials] at some point - even if it's in the supplementary materials."

This is reported in the table on the same page, and also in the R Markdown file.

"Similarly for the handedness check"

Likewise.

"Consider including a similar [individual differences analysis] section for the late ERP component"

We did do this in the R Markdown file, but didn't report it in the main paper. This is because:

1. The results are broadly similar, and we think that the late component here is actually just a continuation of the P3, since the waveforms show that the conditions don't go back to being the same before the late component window. We mention this in the discussion:

However, there may be a simpler explanation: the strong correlation between P3 effect magnitude and late positive complex effect magnitude (r=0.46, p=0.0124) suggests that the two components overlap to the extent that the observed late positive complex in this experiment is just a continuation of the large P3 effect, not a separate component reflecting a separate process.

2. The individual differences section is about teasing apart the difference between a sound-symbolic effect and a learning/memory effect. The late positivity occurs in all kinds of different experiment paradigms, but the P3 is most often related to learning/memory or error/oddball detection. Moreover, as far as we're aware, there are no other sound symbolism studies involving the P3. This is why it is more important to show a sound-symbolic effect for the P3.

Reviewer C

"It would be helpful to the reader if the different stages were signalled more clearly"

Hopefully the improved experiment procedure graph helps with this.

"I am therefore not suggesting that the authors re-run their analysis, but I do think it would be worthwhile to consider this suggestion for future studies."

Thanks for the suggestions, we'll look into it for the next study.

"Did the authors statistically compare the [2AFC] groups on all these factors?"

Not statistically. Mean age in both 2AFC groups was between 21 and 22, there were 4 male participants in the 2AFC high group and 3 male participants in the 2AFC low group, 3 left handers in the 2AFC high group and 2 left handers in the 2AFC low group. Year of post-high school study was varied, with a mean of 2.79 in the 2AFC high group and 2.07 in the 2AFC low group. All participants had learned at least one other language than Dutch (usually English), and all participants except one (in the 2AFC low group) reported speaking at least one other language than Dutch regularly. There was a variety of higher education subjects being studied across the two groups.

Editor Decision for Version 2

Editor: Max Coltheart
Affiliation: Macquaire University, Australia
Editor decision: Revisions required
Decision date: 25 May 2016

Dear Gwilym,

After review, I have reached a decision regarding your submission to Collabra, "How iconicity helps people learn new words: neural correlates and individual differences in sound-symbolic bootstrapping", and am happy to accept your revision for publication, pending the completion of copyediting and formatting processes. Thanks for doing such a responsive and through job in dealing with the reviewer's comments. Very interesting paper, especially for me, who knew nothing about this filed of research

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. The accepted submission will now undergo final copyediting. You will be contacted once this is complete to answer any queries that may have arisen during copyediting and to allow a final chance to edit the files prior to typesetting. If you wish to view your submission during this time, you can log in via the

journal website.

The review information should be included in this email.

Best wishes,

Max Coltheart

Macquarie University

max.coltheart@mq.edu.au
