Protocol for an experimental study of iterated learning and language evolution

Hyoyeon Lee, Seth Bullock and Conor Houghton

Intelligent Systems Laboratory, University of Bristol, United Kingdom
conor.houghton@bristol.ac.uk

Abstract

The iterated learning model simulates the transmission of language from generation to generation. This simulation inspired an experiment in which human participants performed a language learning task similar to the one performed by agents in the simulation. These experiments used a set of 27 images, each image has three properties: shape, colour and movement, with three possible values for each property. A participant is shown a subset of these images, each paired with a word that describes it; after a series of training and testing sequences, a final test elicits the word they believe matches each of the 27 images, a set that includes images they have not been provided a word for in training. These words are then used for training the next participant, allowing the language to evolve along a "chain" of participants. In our experiments we intend to reproduce these original experiments, but using an online platform, and to add a novel experiment which includes a new training and testing protocol which we believe may change the way the language evolves.

Introduction

This is a protocol for a set of three experiments which will be run online using `JSPsych` with participants recruited using `Prolific`. Each experiment will produce a set of chains in which the "language" produced by one participant forms part of the input for the next.

EXPERIMENT 1 and EXPERIMENT 2 are based on the two experiments described in Kirby et al. (2008); our versions differ most obviously in being run online rather than in person. There are other differences. For example, we use animated GIFs whereas the original experiment used still images with arrows to suggest motion. Since we use moving shapes and one movement is spinning in place, we have replaced the circle in the original experiment with a cross.

We have also changed the trial structure in a way that makes the experiment shorter. This will be briefly revisited after the trial structure has been explained. EXPERIMENT 3 is intended to extend the original experiments to include learning trials analogous to the unsupervised learning described in the simulations reported by Bunyan et al. (2025).

An outline of the experiments

The stimuli for these experiments are a set, $\mathcal{C}$, of 27 animated GIFs; for simplicity these will be referred to as images even though "moving images" or "animations" might be more accurate. The images have three properties, movement, color and shape and each property has three values. For movement this is spin, slode and bounce; for colour: red, blue and black; and for shape: square, triangle and cross. These images are available online Lee et al. (2025a).

In the experiment each of these images is paired with a description. This description changes as the experiment continues. The initial descriptions are all single words and for simplicity "word" is used to describe these descriptions even though it is possible for a participant to introduce a space into a description. The mapping of images in $\mathcal{C}$ to words is called a language.

At the onset of each trial a subset $\mathcal{B}_n$ of size n is selected at random from $\mathcal{C}$. For the experiments described here $n = 9$. This is the bottleneck set and in EXPERIMENT 1 and EXPERIMENT 3 these images, along with the words paired to them by the previous participant, are the exemplars used to teach the participant which word goes with which image. At the end of the training protocol, which includes some intermediate testing, the participant is asked to match a word to every image in $\mathcal{C}$; this will include images that were not encountered during training and hence have not had a word associated with them. These images may not have been seen before, or may only have been seen during an intermediate testing phase.

The experiment produces chains in which the word paired to each image in $\mathcal{C}$ by one participant is used when using the bottleneck set $\mathcal{B}_n$ as exemplars for the next participant in the chain. The initial set of words, the ones used for the first participant, are calculated randomly. Each word is a string of two, three or four of the syllables

"da", "vi", "ho", "wi", "nu", "ri", "bi", "ka", "tu"

The length is picked at random from $\{2, 3, 4\}$ and the syllables are picked randomly with replacement.

In all three experiments the chains will be ten trials long

with each trial corresponding to a single participant. The aim is to produce ten chains for each experiment. However, as explained later, there are conditions under which an experiment will be abandoned after five chains or under which a chain will be restarted. Otherwise each chain will be allowed to evolve over all ten participants and each experiment will run until there are ten chains.

In EXPERIMENT 2 the bottleneck set is pruned to avoid duplications. In Experiment 1 in Kirby et al. (2008) and again in our pilot, participants may choose to employ the same word to describe multiple different images causing the language to "collapse" so that by the end of the chain there are only a small number of words and these words appear to describe just one property of the image. EXPERIMENT 2 is intended to avoid this by removing exemplars with duplicate words from the bottleneck set of the next participant. To do this at the start of each trial a new set $\mathcal{P}_n$ is formed. It is identical to $\mathcal{B}_n$ except that it has been pruned so that each included image is paired to a unique word; if two or more images have the same word, one is selected randomly. This means n images are picked from $\mathcal{C}$ to form $\mathcal{B}_n$ and this is then pruned to form $\mathcal{P}_n$, a smaller set. The n indicates that $\mathcal{P}_n$ is formed by pruning $\mathcal{B}_n$, it may have fewer than n elements itself.

There is a danger that $\mathcal{P}_n$ becomes very small; this is a greater potential problem for this experiment than for the original experiment reported in Kirby et al. (2008) because we intend to use a smaller value of n. If only a small number of duplications are produced by each participant it will not be a problem, but if a larger number is, it may be. This is something that our experiment will reveal and, as detailed below, if it does prove a problem the experiment will be halted and the protocol will be supplemented with a new version.

Recruitment

Based on our pilot studies, the experiments are estimated to take 30 minutes to run. Participants will be recruited on `Prolific` using the description:

> In this study we will show you a set of animated figures along with words in an "alien language" and then test you to see how many you remember!

and an anticipated median time of 30 minutes. Participants will be offered £4.50 with a £1 bonus payment for the top 20% performing participants. The difficulty of the task can change from experiment to experiment and as the chain progresses, the "top performers" will be calculated on a per-experiment and per-position-in-chain basis; this detail is not explained to participants.

Recruitment will be limited to United Kingdom-based participants, who state they are fluent in English and have stated "[n]o, I have no issues seeing colours". There are no further demographic constraints, nor do we record any further demographic information. Participants are recruited to at most one trial of these experiments and any potential participant who took part in our pilot experiments is excluded.

Figure 1: This is a still from one example image that might occur in the bottleneck set, B_9, showing the sliding black triangle and matching it to the word "biri"; in the actual animated stimulus the triangle slides horizontally from left to right.

The structure for each trial

Notation

It is easier to describe the trial structure if a notation is first defined to denote each of the stages of the trial.

- B_n: In B_n, images are presented from $\mathcal{B}_n$, each alongside the current paired word. All n images from $\mathcal{B}_n$ are presented. For each presentation the word appears for 1 s and then the word and paired image for 5 s during which the animated movement unfolds. Between successive presentations a fixation cross is displayed for 1 s. The order is random and rerandomised each time.
- P_n: The pruned presentation, P_n, is the same as B_n except that it uses all images from the pruned set $\mathcal{P}_n$. It may included fewer than n items if there are fewer than n unique items available.
- T_{2m}: In the test session, T_{2m}, $2m$ images are presented: m from $\mathcal{B}$ and m from $\mathcal{C} \setminus \mathcal{B}$, each randomly selected in random order. The participant cannot begin typing for the first 2 s that the image is displayed. They cannot progress until they have typed something, so if they have not responded after 5 s the image is displayed again.
- F: This is the full test. It is similar to T_{2m} but in this case all 27 images in $\mathcal{C}$ are tested, in a random order.
- TC_m is a choice test. In the 'T-phase', m images are tested, as in T_{2m}. The subsequent 'C-phase' follows the same order as the T-phase, it is based on the same set of m images in the same order except that each image is presented alongside a distractor image that differs from it in

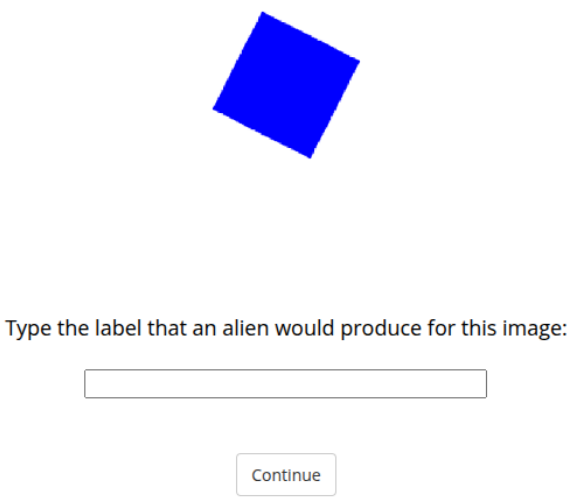

Figure 2: This is a still from one example image that might occur in the test set, T_m, showing the rotating blue square and asking the participant to input the correct word.

one attribute. The word the participant entered in the corresponding T-phase is also provided and the participant is asked to identify which of the two images matches that word.

Structure

The three experiments have the following structure:

- EXPERIMENT 1: $(B_9T_8)^3\ B_9F$
- EXPERIMENT 2: $(P_9T_8)^3\ B_9F$
- EXPERIMENT 3: $B_9(TC_1)^4\ B_9(TC_2)^2\ B_9(TC_4)\ B_9F$

In EXPERIMENT 3 the structure of the intermediate testing phase changes so that this becomes progressively harder: in $(TC_1)^4$ after writing the word the participant is immediately asked to identifying which of two images that word refers to, whereas in (TC_4) four images are tested before the same four images, each paired with a distractor, is used for the choice task. However, in each case four images are used and, for each intermediate test, two of these are from $\mathcal{B}_9$ and two from $\mathcal{C} \setminus \mathcal{B}_9$ with this selection made independently and randomly for each of the three intermediate tests.

By comparison, the previous experiments Kirby et al. (2008) had

- Experiment 1: $(B_{14}B_{14}T_{14})^2\ B_{14}B_{14}F$
- Experiment 2: $(P_{14}P_{14}T_{14})^2\ P_{14}P_{14}F$

Other parts of the trial

In addition to the experimental structure described above, all three experiments have other elements to provide instruction, to obtain consent and so on. A copy of the `JavaScript` programme can be found at Lee et al. (2025b). An overview is provided here.

Participants are asked to enter their Prolific ID. This is followed by a participant information sheet and a consent form. These can be found in the supplementary information. Next there is a set of initial instructions; these instructions are broadly similar to those in the original experiments Kirby et al. (2008). They are split across two screens:

> Welcome to Beta–3–6a in a galaxy far, far away.
>
> We have encountered an intelligent alien life form with its own form of language. You must try to learn this language as best you can. Don't worry if you feel overwhelmed—the alien knows that this is a difficult task for you to master, and that you will make mistakes. It will do its best to understand everything that you say.
>
> (Press ENTER to continue)

and

> You will see a series of pictures and the way in which the alien would describe those pictures. Every now and then the alien will test your knowledge of the language by showing you a picture without any description. Simply write what you think the correct description is in the input box provided.
>
> Don't worry if you feel you have not yet mastered the language! The most important thing is to maintain good relations with the aliens and give it your best shot. ALWAYS GIVE AN ANSWER. That way the aliens will know you are trying. They will go out of their way to try to understand everything you say, and they are very patient.
>
> You will be given a break every 5 minutes or so. Good luck!

There are further, more detailed, instructions. These are given in the supplementary information.

After the experiment, there is a brief questionnaire introduced by

> Questionnaire
>
> You have now completed the experiment and reached the final questionnaire. Please answer the following questions about your background and your experience in this study. Your responses will help us better understand the results.

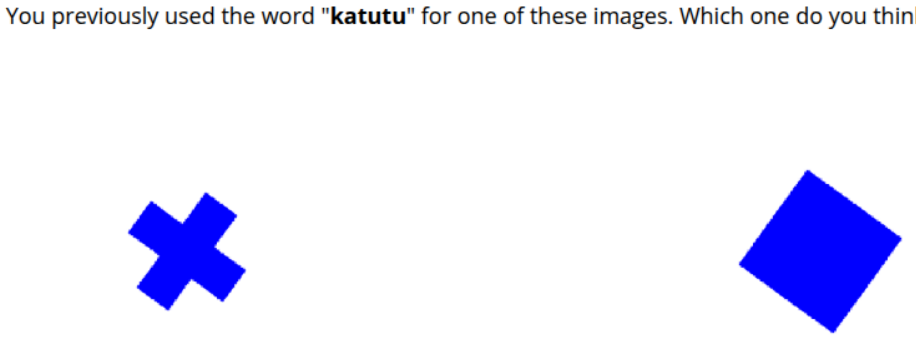

Figure 3: This is a still from one example image that might occur in the choice set, C_m; this is intended to mimic a choice test that might follow the input test in Fig. 2 if the participant had input "katutu" to describe the rotating blue square. In this example, the correct response would be to click on the right panel.

This is followed by six questions divided across three screens. There is a continue button at the bottom of each screen. It is possible to continue without giving any answer.

- What is your first language?
- Have you ever learned any other languages? Please briefly describe.
- What are your thoughts on this study overall?
- Did you notice that some test images were not shown to you during training?
- How do thoughts usually come to you? Do you tend to think more in words, in images, or in some other way? Please specify.
- Please describe any tricks such as mental strategies or written notes you used to assist your memory during the experiment. Do not worry about the answer; your answer will not affect your payment or any potential bonus. Your answer will help us improve our study.

The questionnaire is followed by a debrief page outlining the motivation for the experiment. The text is given in the supplementary information. A final "Finish Experiment" button redirects participants back to Prolific.

Data analysis and presentation

There is no intention to do formal hypothesis testing with the data that will be collected in these three experiments. This is a qualitative study. All the data will be made available on `osf.io`, `zenodo.org` or another non-commercial data repository. It is intended to write a summary of the results, either for publication or to place on an archive. Largely, this will include illustrative examples of languages from the chains. We will also include summary statistics, such as the average number of words at different stages in the chain. For comparison with Kirby et al. (2008) we will calculate test scores, again, at different stages of the chain.

One central point of interest is the compositionality of any language: the extent to which a language features words for which different parts of a word refer consistently to different properties of the associated image. We will discuss this, illustrating either its occurrence or absence with examples and, possibly, with summary statistics. We will also consider stability, the degree to which a language remains the same from participant to participant; this will also be illustrated using examples and, possibly, summary statistics.

If, as expected, the language in EXPERIMENT 1 partially collapses, we will summarise which image properties, if any, are encoded in any partially collapsed language; it will be interesting if, for example, the "verb-like" property wins out over the "noun-like" or "adjective-like" properties, for example.

Criteria for finishing experiments early

We intend to run the experiment until we have ten chains each comprising ten consecutive participants for all three experiments. However, we will stop if there are obvious problems with the `JSPsych` code that runs the experiment.

We will also consider halting EXPERIMENT 3 if the average number of words at the end of the chain is less than six; EXPERIMENT 3 is intended to examine whether the reflective design of the test stages prevents the number of words from falling as it did in Experiment 1 in Kirby et al.

(2008). If this is not working we will consider halting the experiment, adjusting the structure and running the experiment again; this will be described in a supplementary protocol; this supplementary protocol will make it clear that it includes a revised protocol.

If there is evidence EXPERIMENT 2 is not working because the size of $\mathcal{P}_n$ falls to four or fewer image-word pairs for at least one chain in the first five chains, this experiment will be halted. A new version will be run which ensures that this does not happen. This will involve selecting further examples from $\mathcal{C}$ or even introducing new random words. In this case the new design will be described in a supplementary protocol.

In assessing stability, it may become obvious that longer chains are needed. If that is the case we intend to produce a supplementary protocol extending some subset of the existing chains to 15 or 20 trials.

If the estimate of the experimental run time proves to be very wrong, particularly if the estimate is too short, we will change the estimate and offered a new reward based on a rate of £9 for an hour.

In our pilot we saw one chain where one participant wrote English descriptions of each image. In the planned experiment any chain where this happens will be halted and restarted using the words from the previous participant.

Participants can withdraw, in this case their partial data will be included in the dataset but not used for any summary statistics. The chain will be restarted at the previous participant.

Code and data availability

The page `iteratedlm.github.io` acts a general gateway to this project and other projects related to language learning. This will link to any data repository and to the code used to run the experiments and analyse the data. The two `github` repositories:

- `IteratedLM/2025_06_animations`
- `IteratedLM/2025_06_human`

Lee et al. (2025a) and Lee et al. (2025b) are archived versions of these two repositories.

of the code used to make and run the experiments are

Ethical approval, funding and acknowledgements

This study was approved by the Faculty of Engineering Research Ethics Committee at the University of Bristol (Ref: 26048).

The project was supported by a Leverhulme Research Fellow (RF-2021-533 to CH) and by UKRI grant EP/Y028392/1: AI for Collective Intelligence (AI4CI).

Thank you to Hannah Cornish for providing us with the stimuli used in the original experiments and for suggesting that we use animated GIFs in our experiments. We have not reused the original stimuli, but it was very useful to see them.

References

Bunyan, J., Bullock, S., and Houghton, C. (2025). An iterated learning model of language change that mixes supervised and unsupervised learning. *PLOS Complex Systems*, 2(3):e0000030.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Lee, H., Bullock, S., and Houghton, C. (2025a). IteratedLM/2025_06_animations: Stimuli for an experimental study of iterated learning and language evolution (v1.0).

Lee, H., Bullock, S., and Houghton, C. (2025b). IteratedLM/2025_06_human: Code for an experimental study of iterated learning and language evolution (v1.0).