The Picture Guessing Game: The Role of Feedback in Active Artificial Language Learning

Felicity F. Frinsel (fff26@cornell.edu)
Department of Psychology, Cornell University, Ithaca, New York 14853

Fabio Trecca (fabio@cc.au.dk)
School of Communication and Culture, Aarhus University, 8000 Aarhus, Denmark

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, New York 14853
School of Communication and Culture, Aarhus University, 8000 Aarhus, Denmark

Abstract

Language is acquired within a complex, interactive environment. A key question for cognitive science is whether and how different types of environmental cues might affect the learning and processing of language. In this paper, we explore the role of feedback as a possible cue in a novel active artificial language learning task: The Picture Guessing Game. Subjects were instructed to guess which scene correctly displayed the meaning of a spoken sequence of unfamiliar monosyllabic words. After their response, either positive, negative, or no feedback was provided. The prediction was that feedback would help the subject to eventually learn the vocabulary, syntax, and semantics of the artificial language. The results indeed showed that feedback (both positive and negative) is beneficial and necessary to attain a certain level of learning. Interestingly, the data showed that positive feedback may be particularly helpful for the learner, promoting more in-depth learning of the artificial language.

Keywords: artificial language learning; feedback; language acquisition; multiple-cue integration

Introduction

Statistical learning (SL), a domain-general learning mechanism that enables individuals to utilize distributional properties of sensory input in order to learn probabilistic regularities, has become a foundational element in cognitive science (see Armstrong, Frost, & Christiansen, 2017, for a review). Although the first artificial language learning study was conducted almost a century ago (Esper, 1925), research on SL in the context of language acquisition and processing has expanded significantly after the seminal study by Saffran, Aslin, and Newport (1996), showing that infants are sensitive to the transitional probabilities of syllables. Following this work, a vast number of studies have reported humans' ability to detect patterns in artificial and real-world input starting from a very young age by using statistical cues without any explicit feedback (e.g., Maye, Werker, & Gerken, 2002; Saffran, Aslin, & Newport, 1996; Saffran, 2003). In order to study SL, several fairly simple paradigms have been used, such as the widely adopted artificial language learning (ALL) paradigm. This paradigm opens up the possibility to investigate language learning abilities in a controlled environment, as the artificial language permits the researcher to control the learner's input. The methodological nature of such paradigms has recently been challenged, however (e.g., Armstrong et al., 2017; Frost, Armstrong, & Christiansen, 2019). For example, most of these paradigms involve passive exposure to recurrent patterns (Christiansen, 2019), often with separate learning and test phases. Yet, such passive exposure, followed by testing, provides little information about what is driving the learning process and how it develops across time. Moreover, important aspects of language learning in a natural environment, such as its interactive nature and the integration of multiple cues, are often also not considered in such passive exposure paradigms.

Given that language is acquired in a complex and noisy context, learning and processing a language requires successfully integrating multiple cues. For instance, learners must successfully integrate syntactic and semantic cues to the meaning of an utterance in order to learn and process language (e.g., Gibson, Bergen, & Piantadosi, 2013). Additionally, language is immersed in a rich and dynamic environment in which social interactions seem to play an important role in the acquisition of language (Elmlinger, Schwade, & Goldstein, 2019; Goldstein & Schwade, 2008; Romeo et al., 2018). Yet, the input used in ALL experiments are generally overly simplified, isolated, and confined as opposed to linguistic input in the real-world. Even though previous ALL studies have made major contributions to the field, in order to gain further insight into language acquisition, novel ALL paradigms that simulate a more naturalistic environment in which learners acquire structures in a meaningful and interactive context are necessary (see Frost et al., 2019).

We therefore developed and tested a novel experimental paradigm: *The Picture Guessing Game*. Crucially, this paradigm allows for the study of active statistical learning in light of multiple-cue integration using constructions that are more language-like (see Method section for more details). In this paper, we present this new artificial language learning paradigm, in which the learner is not just a passive participant but instead actively makes guesses as part of the learning process. These responses allow us to gain insight into the trajectory of learning. The paradigm additionally permits us to explore both the individual and interactive effects of multiple explicit and implicit cues on learning. Here, we used this paradigm to explore the role of feedback (either positive or negative) on the learning of regularities relating to syntactic and semantic information in the speech input.

Feedback in Language

In the field of language acquisition, the existence of feedback has a long controversial history, going back more than half a century (Schachter, 1991). Gold's (1967) theorem showed that only finite-state languages could be acquired from positive evidence. To learn more complex languages—context-free and beyond—required additional constraints on learning. One option was to hypothesize the existence of built-in biological constraints (such as an innate Universal Grammar, UG; Chomsky, 1965). Another possibility was if the child could receive negative feedback — being told explicitly every time they produced an ungrammatical utterance. The idea of negative feedback, however, conflicted with the longstanding belief that children do not receive, need, or use any (corrective) feedback in order to learn a language (e.g., Marcus, 1993; Ramscar & Yarlett, 2007). For that reason, the UG hypothesis was often considered the only viable option. Further supporting the conclusion, Baker (1979) used the 'no negative evidence problem' to support the idea that children would exclusively use positive evidence to rectify incorrect suppositions. A paradox emerged here (sometimes referred to as Baker's Paradox), as the question of how learners could recover from overgeneralization without having any negative feedback available to them stayed unanswered. This, in combination with later theoretical changes and developments in linguistics, affected the research directions within language acquisition extensively. In fact, the focus has mainly been on providing evidence showing that negative feedback does exist.

Although there is substantial evidence against the idea that children receive or use explicit negative evidence, other studies have revealed that other types of feedback, typically provided implicitly, are available for language learning (Chouinard & Clark, 2003; Saxton, 2000). Interestingly, even though the possibility of *positive* feedback has been acknowledged within the study of language acquisition, the focus has been almost entirely on negative feedback. Accordingly, empirical data on, for example, the exact role of positive feedback in language learning is sparse. It is therefore not surprising that the role of feedback has largely been neglected within ALL (but see Dale & Christiansen, 2004), even though this paradigm has been used extensively to explore learnability issues in language learning.

The objective of this study was therefore to examine the effect of both positive and negative feedback on learning while simultaneously integrating syntactic and semantic information, using the Picture Guessing Game designed to model the acquisition of language structures under more complex circumstances. To do so, three feedback conditions (positive feedback, negative feedback, no feedback) were implemented. The conditions in which learners received feedback were intended as initial steps toward incorporating social interactions into artificial language learning context. We hypothesized that both positive and negative feedback would facilitate the learning of the artificial language compared to no feedback. Moreover, we predicted that positive feedback would provide for better learning than negative feedback (pre-registration: https://aspredicted.org/8dk2b.pdf).

Method

Participants

One hundred and twenty Cornell University undergraduates (84 females; age: M = 19.7, SD = 1.6) participated in exchange for course credit. All subjects were native English speakers. Subjects were randomly assigned to one of the three feedback conditions: *Positive Feedback* (N = 41), *Negative Feedback* (N = 39), and *No Feedback* (N = 40).

Materials

Auditory Stimuli The artificial language used in this study consisted of twelve monosyllabic nonsense words inspired by Dale & Christiansen (2004). Nine of the words (*hep*, *jove*, *rus*, *lem*, *kav*, *rud*, *pel*, *hef*, *jux*) were used as nouns, and three (*poox*, *sook*, *voop*) served as verbs. Each noun was randomly paired with a unique animate (human or animal) referent on screen, and each verb was assigned a unique arrow shape, resulting in twelve distinct sound-symbol pairings. A total of 240 spoken sequences were generated using a speech synthesizer, each consisting of a verb and three distinctive characters: an agent, an object, and a recipient. The sequences of nonsense words used a non-English SOV word order to avoid any facilitation from the subjects' native language and followed one of two dative structures: 1) a prepositional dative (PO), with structure S-O-prep-R-V; e.g., *rud hef ma-jove poox* (the clown a monkey to the girl shows); and 2) a double object dative (DO), with structure S-R-O-V; e.g., *rud jove hef poox* (the clown the girl a monkey shows).

All sentences in the artificial language were semantically plausible, with human characters in the role of agents (S) and recipients (R), and animal characters in the role of objects (O). However, during the test phase (see Procedure section), the meaning of some of the sentences was manipulated by reversing the thematic roles between agent and object (ImplausibleS; e.g., *the monkey a clown to the girl shows*) or between object and recipient (ImplausibleR; e.g., *the clown a girl to the monkey shows*), resulting in semantically implausible (albeit still semantically possible) events as compared to real-world semantics.

Visual Stimuli For each spoken sequence, four scenes were depicted on the screen. Each scene illustrated the thematic relations between the same four constituent elements (S, O, R, and V). However, only one scene matched the aurally-presented target sentence. Figure 1 shows an example trial, in which the correct match for the target sentence is the scene in the upper-left corner. The other three pictures are foils corresponding to incorrect interpretations of the target sentence. To exclude the possibility that the participants would select the correct picture by solely relying on the arrow-verb mapping (which corresponded to the last word in the sentence), the same arrow was depicted in all four pictures. This made it

impossible to disambiguate the correct picture without taking the verb argument structure into account.

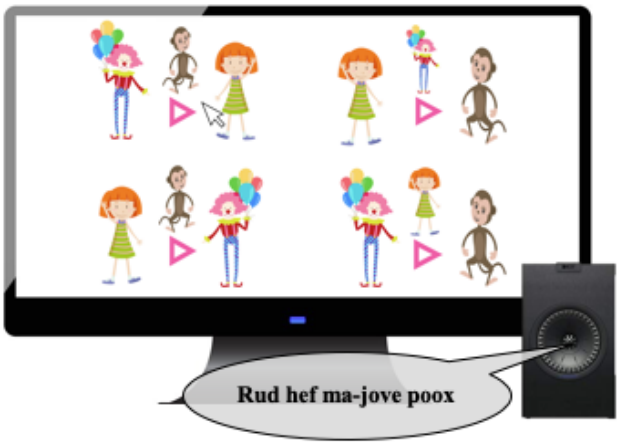

Figure 1: Example stimuli training and test phase

Procedure

The experiment was run using PsychoPy2 version 1.90.3 (Peirce & MacAskill, 2018). Subjects were seated in front of a computer screen and wore headphones during the course of the experiment. The study consisted of four parts: a learning phase, a test phase, a verb test, and a noun test. All subjects completed the experiment in this exact order, but the order of presentation of individual sentences within each phase was fully randomized across participants. The procedure of the learning and test phase were identical and for that reason there was no noticeable distinction between the two from the perspective of the learner. The robustness of learning, however, was tested by manipulating semantic plausibility in the test phase only, by introducing ImplausibleS and ImplausibleR sentences.

Before each trial, a fixation cross was shown at the center of the screen for one second. Subsequently, subjects were presented with four scenes located at the corners of the screen while listening to the spoken sequence (Fig. 1). They were instructed to click on the scene corresponding to their interpretation of the sentence they had just heard. For each trial, the mouse cursor was automatically repositioned at the center. The location of the four picture choices was fully randomized across trials and subjects. The learning phase consisted of five blocks of 40 trials and the test phase comprised one block of 40 trials. After completing block two and four, subjects were informed about their progress into the experiment, while given the option to take a short break. After completing the test phase, nine additional sequences were presented during a verb test in order to determine whether subjects successfully learned the verb labels. The stimuli were presented in the exact same manner as during the training phase, but rather than showing the same arrow in all four scenes, the visual symbol for the verb varied for three of the four scenes. Following the verb test, subjects completed a noun test, in order to ascertain that they successfully learned all sound-symbol pairings. Here, for each trial an animate referent was presented individually at each corner of the screen while listening to words presented one at a time (see Fig. 2 for an example). The experiment took approximately one hour to complete.

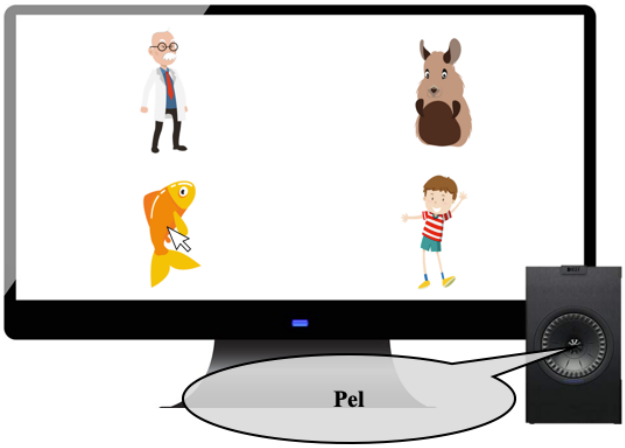

Figure 2: Example stimuli noun test

Subjects allocated to the Positive Feedback condition received feedback both auditorily and visually after correctly selecting the target scene: the spoken sequence was repeated while the foil images turned white, leaving only the correct picture on the screen. Subjects assigned to the Negative Feedback condition were simultaneously exposed to the correct spoken sequence and a red X across the screen after selecting a foil. No form of explicit feedback was given after selecting the correct scene. Thus, even though subjects in both feedback conditions got exposed to the correct sentences, subjects in the Negative Feedback condition only got information about which scene was incorrect (i.e., the one they selected), but no information was provided about the correctness of the other three scenes. In the No Feedback condition, no feedback was given, and the auditory sequence was repeated after each mouse click independent of the accuracy of the subject's response. Lack of learning in the No Feedback condition could therefore not be the result of less exposure to the artificial language. Subjects did not receive any specific instructions about the feedback in either of the three conditions. This approximates the nature of language learning in the real world, where the role of feedback is implicit and has to be learned.

Data Analysis

The data was analyzed using generalized linear mixed-effects models using the packages lme4 version 1.1-21 (Bates, Mächler, Bolker, & Walker, 2015), car 3.0-3 (Fox & Weisberg, 2019), emmeans 1.4 (Lenth, 2019) and lmerTest 3.1-0 (Kuznetsova, Brockhoff, & Christensen, 2017) in *R* version 3.6.1 (R Core Team, 2019) and RStudio 1.0.153. All models contained by-subject and by-item random intercepts. In order to calculate main and interaction effects, Type II Wald Chi-square tests were run. The *emmeans* function was applied to detect significant differences between contrasts. Our hypotheses and to be conducted analyses were pre-registered on AsPredicted.

Results

Training Phase

A generalized linear mixed-effects model with a logit link function was fitted in order to test accuracy in picture choice. We found a main effect of Feedback ($\chi^2(2) = 75.29$, $p < .001$), a main effect of Syntax ($\chi^2(1) = 37.87$, $p < .001$), and an interaction effect between Feedback and Syntax ($\chi^2(2) = 95.11$, $p < .001$). Separate tests for each level of Feedback showed a significant Block effect within the Positive ($\chi^2(4) = 195.66$, $p < .001$) and Negative ($\chi^2(4) = 72.54$, $p < .001$) Feedback condition, indicating that in both feedback conditions, subjects' performance significantly increased over time. In the No Feedback condition, subjects, on average, only selected the correct scene 20% of the time. Hence, as shown in Figure 3, no learning was observed within the No Feedback condition ($p = .55$). Further analyses by means of contrast comparisons indicated that subjects' accuracy in the Positive Feedback condition did not differ significantly from those in the Negative Feedback condition ($\beta = -0.04$, $SE = 0.20$, $z = -0.18$, $p = .98$): Subjects in the Positive Feedback condition selected the correct scene equally often on average (50%) as subjects in the Negative Feedback condition (49%). Consequently, the amount of exposure to feedback, and thus the number of times the subject heard the sequence twice, was similar for both feedback conditions.

Moreover, a significant effect of Syntax was observed in the No Feedback condition ($\chi^2(1) = 165.98$, $p < .001$) as well as in the Positive Feedback condition ($\chi^2(1) = 8.94$, $p = .003$). Put differently, significantly more correct responses were given on PO structures compared to DO structures, paralleling sentence processing results in natural language (Gibson et al., 2013). Note though that the significant effect of Syntax in the No Feedback condition may be spurious, as the percentage of correct responses for PO and DO structures was at or below chance. Interestingly, no such significant effect was found in the Negative Feedback condition ($p = .084$).

Test Phase

There was a main effect of Feedback ($\chi^2(2) = 14.18$, $p < .001$), a main effect of Syntax ($\chi^2(1) = 33.48$, $p < .001$), a main effect of Plausibility ($\chi^2(2) = 84.19$, $p < .001$), and an interaction effect between Feedback and Syntax ($\chi^2(2) = 53.49$, $p < .001$) as well as between Feedback and Plausibility ($\chi^2(4) = 294.91$, $p < .001$). Follow-up analyses of Feedback within the factors Syntax and Plausibility were performed by means of contrast comparisons. These analyses revealed that the main effects and interactions were mainly driven by patterns in the No Feedback condition that differed from those observed in the other two feedback conditions.

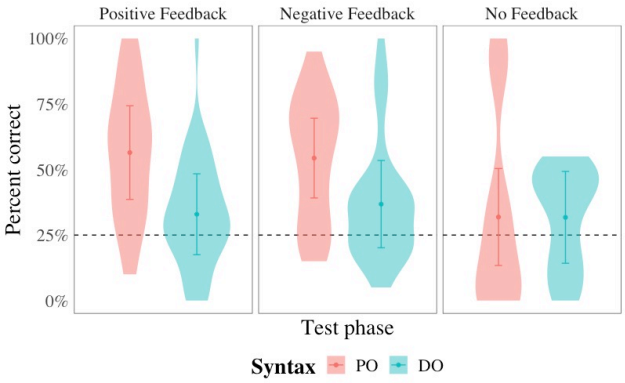

Figure 4: Performance on PO and DO structures split by feedback condition

More specifically, as shown in Figure 4, significantly more correct responses were given on PO compared to DO structures in both the Positive ($\beta = -1.11$, $SE = 0.14$, $z = -7.68$, $p < .001$) and Negative Feedback condition ($\beta = -0.85$, $SE = 0.15$, $z = -5.78$, $p < .001$); whereas no difference in performance was observed between PO and DO structures when no feedback was provided ($\beta = -0.01$, $SE = 0.15$, $z = -0.05$, $p = .96$). Accordingly, only performance on PO sequences in the No Feedback condition was significantly worse as compared to the

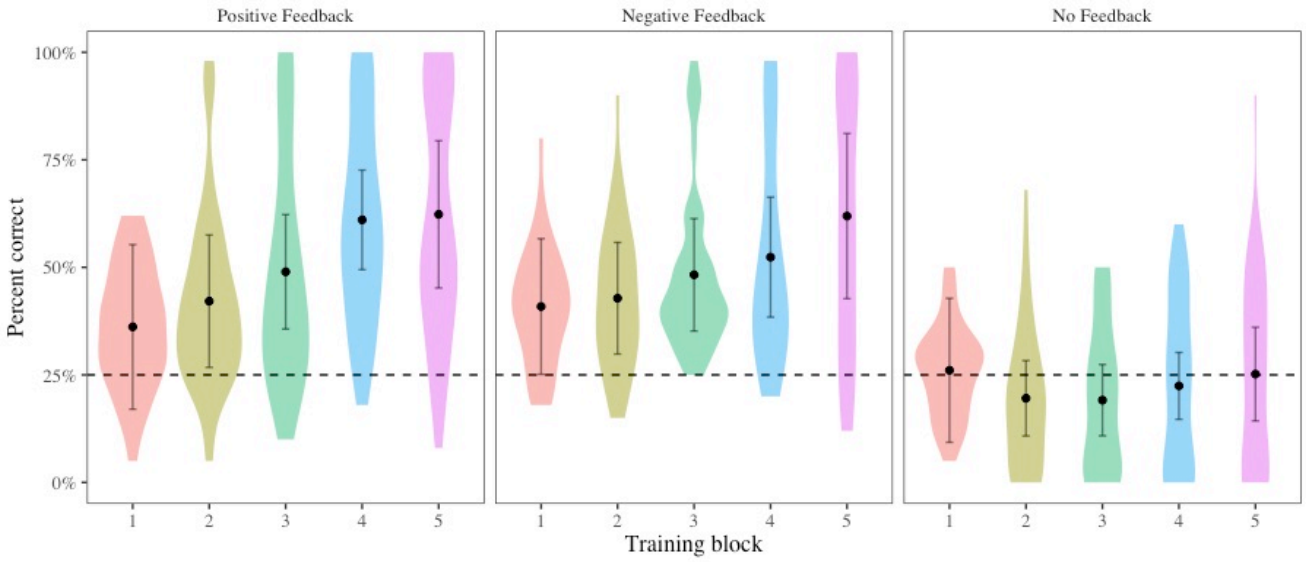

Figure 3: Learning over training blocks split by feedback condition

performance in the Positive (β = −1.13, *SE* = 0.20, *z* = −5.61, *p* < .001) and Negative (β = −1.05, *SE* = 0.20, *z* = −5.13, *p* < .001) Feedback conditions. No significant differences were observed when contrasting the accuracy scores in the Positive Feedback condition with those in the Negative Feedback condition (PO, β = −0.08, *SE* = 0.20, *z* = −0.41, *p* = .998; DO, β = 0.18, *SE* = 0.21, *z* = 0.87, *p* = .954).

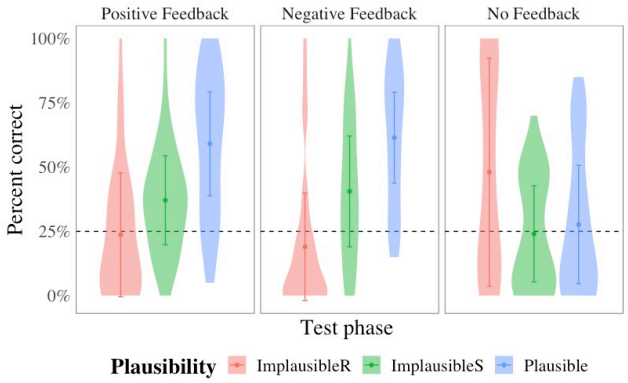

Figure 5: Semantic plausibility performance split by feedback condition

Additionally, as shown in Figure 5, whenever some form of feedback was provided (positive or negative), sentences with plausible semantics led to significantly higher accuracy than sentences with implausible semantics (Positive, β = 1.36, *SE* = 0.12, *z* = 11.08 *p* < .001; Negative, β = 1.58, *SE* = 0.13, *z* = 12.14, *p* < .001). The opposite pattern was found, however, when subjects were withheld from any feedback (β = −0.42, *SE* = 0.12, *z* = −3.42, *p* < .001). Likewise, implausible sentences with animal referents as subjects (ImplausibleS) led to higher accuracy scores than those in with animal referents as recipients (ImplausibleR), only when feedback was provided (Positive, β = 0.73, *SE* = 0.18, *z* = 4.10, *p* < .001; Negative, β = 1.28, *SE* = 0.19, *z* = 6.60, *p* < .001). Again, the opposite pattern was found when no feedback was provided (β = −1.16, *SE* = 0.17, *z* = −6.72, *p* < .001). In fact, significant differences were revealed for all three plausibility levels after contrasting the No Feedback and Positive Feedback condition (Plausible, β = −1.47, *SE* = 0.22, *z* = −6.82, *p* < .001; ImplausibleS, β = −0.66, *SE* = 0.25, *z* = −2.67, *p* = .021; ImplausibleR, β = 1.23, *SE* = 0.25, *z* = 5.03, *p* < .001) and the No Feedback and Negative Feedback condition (Plausible, β = −1.63, *SE* = 0.22, *z* = −7.39, *p* < .001; ImplausibleS, β = −0.81, *SE* = 0.25, *z* = −3.23, *p* = .004; ImplausibleR, β = 1.64, *SE* = 0.26, *z* = 6.36, *p* < .001). Note that the somewhat higher accuracy on ImplausibleS sentences as compared to ImplausibleR sentences may be caused by the fact that non-human characters never occurred in the subject position for plausible sentences. However, the locations of the characters in ImplausibleR-PO sentences (e.g., *the clown a girl to the monkey shows*) were identical to those in plausible DO structures (e.g., *the clown a girl a monkey shows*), but with a preposition added. Similarly, the characters' positions in ImplausibleR-DO sentences (e.g., *the clown a monkey a girl shows*) were identical to those in plausible PO structures (e.g., *the clown a monkey to the girl shows*), but without the preposition. Further inspection of the data revealed that subjects were most likely to select the picture corresponding to plausible structures when presented with ImplausibleR sentences in the Positive (β = 1.05, *SE* = 0.242, *t* = 4.16, *p* < .001) and Negative Feedback condition (β = 0.82, *SE* = 0.24, *t* = 3.39, *p* < .001).

Verb Test

Subjects in all three feedback conditions performed relatively poorly on the verb test. In the Positive Feedback condition, on average, subjects chose the correct picture 34 percent of the time (within-subject *SD* = 0.18), whereas subjects in the Negative and No Feedback condition had an accuracy score of 27% (within-subject *SD* = 0.20 and 0.18, respectively). No significant main effect of Feedback ($\chi^2(2)$ = 4.39, *p* = .111) was detected, however. Contrast comparisons showed that the difference between the Positive Feedback and the two other feedback conditions was only marginally significant (Positive vs. Negative: β = −0.37, *SE* = 0.21, *z* = −1.82, *p* = .068; Positive vs. No: β = −0.36, *SE* = 0.20, *z* = −1.77, *p* = .07).

Interestingly, as shown in Table 1 below, all three verbs were learned significantly above chance and equally well within the Positive Feedback condition (indexed with asterisks). For the Negative Feedback condition as well as the No Feedback condition, variation between some of the verbs, if not all, was observed and none of the verbs were significantly learned above chance (.236 ≥ *p* ≤ 1).

Table 1: Mean (SD) percent correct per verb split by feedback condition.

		Feedback condition		
		Positive	*Negative*	*No*
Verb	*Sook*	34* (.29)	28 (.29)	30 (.31)
	Poox	35* (.30)	25 (.29)	25 (.22)
	Voop	34* (.27)	28 (.31)	27 (.25)

Note: * = *p* < .04

Noun Test

The overall accuracy on the noun test was relatively high for all three feedback conditions. In the Positive Feedback condition, on average, subjects chose the correct picture 79 percent of the time (within-subject *SD* = 0.23), whereas subjects in the Negative and No Feedback condition had an accuracy score of 60% (within-subject *SD* = 0.30) and 64% (within-subject *SD* = 0.27), respectively. All nine nouns were learned above chance for all feedback conditions. However, although no significant difference was found between the Negative and No Feedback conditions (β = 0.23, *SE* = 0.36, *z* = 0.62, *p* = .807), subjects in the Positive Feedback condition learned the nouns better than subjects in the Negative Feedback condition (β = −1.18, *SE* = 0.37, *z* = −3.16, *p* = .005) and subjects in the No Feedback condition (β = −0.96, *SE* = 0.37, *z* = −2.57, *p* = .027).

Discussion and Conclusions

To our knowledge, this is the first study that compared the role of both positive and negative feedback by means of an active artificial language learning paradigm. The data illustrates that feedback is an efficient cue used by learners, and enables them to pick up on syntactic complexity effects, with PO being easier than DO, previously only observed in natural language contexts (e.g., Gibson et al., 2013). Specifically, the overall performance of subjects plotted over training blocks shows a pattern of learning only when feedback was provided. Interestingly, no learning was observed in the No Feedback condition, even though subjects on average got more exposure to the spoken sequences. Possibly due to the active nature of the paradigm and the integration of multiple-cues, simple exposure to the artificial language seems insufficient to fully learn it. Nonetheless, in keeping with findings by Jeuniaux, Dale, and Louwerse (2009), subjects in the No Feedback condition still revealed their ability to pick up on simple statistical patterns at the word level (as indicated by the high performance on the noun test). Moreover, no differences were found between the Positive and Negative Feedback condition regarding the ability to detect and learn syntactic structures and the way subjects responded towards semantically implausible scenes. Interestingly, however, positive feedback seems to confer an advantage over negative feedback, as more robust and divergent learning was observed when subjects were given positive feedback rather than negative or no feedback. That is, verbs and noun were learned better when positive feedback was provided. Thus, the results are in line with our pre-registered predictions.

Although this study was conducted with college-aged subjects who already fully acquired their native language, our findings are encouraging, because the positive feedback condition bears some resemblance to a parental behavior that has mostly been unattended within the field of language acquisition where the focus has been predominantly on negative feedback instead (e.g., Chouinard & Clark, 2003; Clark & De Marneffe, 2012; Lustigman & Clark, 2019). However, despite the fact that the Picture Guessing Game provides a promising new way to explore the role of feedback in ALL, it is unlikely that language learners exclusively receive one type of feedback one hundred percent of the time. Therefore, experiments are currently being conducted to address this issue by incorporating probabilistic feedback (rather than deterministic, as in the version of the study presented here). Such follow-up experiments might give fruitful insights into the amount of feedback that is necessary in order to successfully acquire a language, which subsequently provides information about the effectiveness of those types of feedback. Furthermore, in the current study, the effects of both negative and positive feedback on learning were accessed separately, as it would have been problematic to determine whether their impact on learning differs if both were incorporated simultaneously. Future experiments, however, could explore the combinatory effects of positive and negative feedback in language learning.

Ultimately, these findings support the idea that paradigms simulating a more naturalistic learning environment are necessary in order to obtain more fine-grained information about language acquisition (Frost et al., 2019). The Picture Guessing Game promises to provide a way to experimentally investigate language learning in a more complex and interactive environment. The learner is actively making guesses as part of the learning process, which allows us to gain insight into the trajectory of learning. Note, however, that in its current form, the Picture Guessing Game is still an offline task, but could very easily be transformed into an online task by using mouse or eye tracking in order to obtain even more information about the trajectory of learning. Furthermore, although the focus of this study has been on feedback and its effect on language learning, many other aspects could be explored in future experiments. Additionally, this current study was conducted with college-aged subjects, but the paradigm could easily be modified for child research. All things considered, both the findings from this study as well as the paradigm itself have important implications and contribute novel developments to the field of SL.

Acknowledgement

We thank Phoebe Ilevbare, Eleni Kohilakis, Sophia Zhang, Gauri Binoy, Linda Webster, Emma Goldenthal and Susanne Ruckelshausen for their help with data collection. This research was supported in part by a seed grant from the Interacting Minds Centre at Aarhus University.

References

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: Past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*, 20160047.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533-581.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637-669.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468-481.

Clark, E. V., & De Marneffe, M. C. (2012). Constructing verb paradigms in French: adult construals and emerging grammatical contrasts. *Morphology*, *22*(1), 89-120.

Dale, R., & Christiansen, M. H. (2004). Active and passive statistical learning: Exploring the role of feedback in artificial grammar learning and language. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 262-267). Mahwah, NJ: Lawrence Erlbaum.

Elmlinger, S. L., Schwade, J. A., & Goldstein, M. H. (2019). The ecology of prelinguistic vocal learning: parents

simplify the structure of their speech in response to babbling. *Journal of child language*, *46*(5), 998-1011.

Esper, E. A. (1925). *A technique for the experimental investigation of associative interference in artificial linguistic material*. Philadelphia, PA: Linguistic Society of America.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*. Third Edition. Thousand Oaks CA: Sage

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128-1153.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051-8056.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, *10*(5), 447-474.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515-523.

Jeuniaux, P., Dale, R., & Louwerse, M. (2009). The Role of Feedback in Learning Form-Meaning Mappings. In *Proceedings of the 31th Annual Meeting of the Cognitive Science Society* (pp. 1488-1493). Red Hook, NY: Curran Associates, Inc.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Lenth, R. (2019). emmeans: estimated marginal means, aka least-squares means. R package v. 1.3. 4.

Lustigman, L., & Clark, E. V. (2019). Exposure and feedback in language acquisition: adult construals of children's early verb-form use in Hebrew. *Journal of child language*, *46*(2), 241-264.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53-85.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101-111.

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. London: SAGE.

R Core Team (2019). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927-960.

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language related brain function. *Psychological Science*, *29*(5), 700-710.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110-114.

Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, *20*(60), 221-252.

Schachter, J. (1991). Corrective feedback in historical perspective. *Interlanguage studies bulletin, 7*(2), 89-102.