Neural Tracking of Implicit vs Explicit Phonotactic Learning

Enes Avcu, Ryan Rhodes and Arild Hestvik

University of Delaware

Summary. Artificial grammar learning (AGL) studies have been widely used for testing the learnability of phonological patterns. It has been shown at the behavioral level that learners can extract adjacent and non-adjacent dependencies with relatively short training (Finley, 2017). Less is known about how lab-learned patterns are encoded at the neurophysiological level (cf. Domahs et al., 2009; Moore-Cantwell et al., 2018). The aim of the current study was to examine the neurophysiological correlates of implicit and explicit learning of a non-adjacent phonotactic pattern. The "implicit" group merely repeated grammatical exemplars without any explicit instruction, while the "explicit" group had the rule explained. While recording EEG, participants were presented with words that were either well-formed or ill-formed according to the rule. We found that both groups performed behaviorally with accuracy levels indicating knowledge of the rule. However, only the implicit learning group exhibited an ERP response modulated by well-formedness, which we interpret as reflecting prediction errors. These results show that implicit, but not explicit, learning engages neurophysiological mechanisms that lead to prediction models at the neural level and suggests that implicit lab-learning experiments tap into unconscious, automatic learning that is characteristic of natural language acquisition.

Methods. We ran an artificial grammar learning experiment with two learning conditions (implicit vs explicit), testing the learnability of a simple phonotactic pattern – a non-adjacent sibilant harmony pattern attested in Navajo.

Stimuli. All training and test stimuli consisted of two syllables of the form of CV.CV, with sibilants ([s, ʃ]) as the first and second consonants. All words were either "harmonic" (both sibilants identical) or "disharmonic" (mixed [s] and [ʃ]). The duration of each phoneme was strictly controlled at 100ms, making each word 400ms long, and the violation at 200ms.

Procedure. 45 monolingual American English speakers participated, divided into two groups (N=24 and 21). The procedure for the implicit-learning group consisted of two phases: training and testing. The training phase differed for the two groups. For the explicit-learning group, the rule was explained: "s and ʃ cannot appear in the same word". Explicit-learning participants were then presented with all the harmonic and disharmonic words and instructed to press a button in response to each stimulus to categorize them. Feedback was given for correct and incorrect responses. Implicit-learning participants instead listened only to harmonic words and were instructed to repeat each word orally. Implicit-learning participants were not told the rule and received no feedback. In the testing phase, participants from both groups were instructed to listen to a sequence of words and categorize each word as "part of the language" (i.e. novel harmonic words) or "not part of the language" (novel disharmonic words) that they had been exposed to during training. Participants were tested in an auditory oddball paradigm, with 80% harmonic words and 20% disharmonic words. The groups differed only in training (explicit vs implicit)[1].

Data Recording and Analysis. *Hits* (a disharmonic word was presented, and the participant reported it as disharmonic) and *false alarms* (a harmonic word was presented, but the participant reported it as a disharmonic word) were used to calculate *d'*, a measure of the participant's sensitivity to the rule. Learning was then modeled as having a *d'* greater than 0. EEG was recorded with a HydroCel 128 electrode net (Electrical Geodesics). The P3 measurements were taken from the rare-minus-frequent difference waves measured at frontal (F3, Fz, F4), central (C3, Cz, C4), and parietal (P3, Pz, P4) electrode sites. ERPs were computed for two-time windows: a stimulus-locked P3 (400 to 700ms after the stimulus onset), and a response-locked P3 (-200 to -100ms before the behavioral response), following Luck

[1] Testing phase was the same for both groups, except the implicit group had 300 trials compared to 1200 in explicit group.

(2009). Analysis of variance (ANOVA) included factors of the region (frontal, central, parietal) and harmony (harmonic, disharmonic words).

Results. Behavioral results showed that the implicit group detected disharmonic words with a mean sensitivity (d') of 0.558, while the explicit group's sensitivity was 1.666. The difference between groups was significant: $t(43)=3.68$, $p<.001$, $1-\beta=.976$. The implicit group's mean accuracy was 0.66 (SD=.13), while explicit group's mean accuracy was 0.80 (SD=.14), which also showed a significant group difference ($p=.002$, $1-\beta=.949$). Electrophysiological results for the stim-locked and resp-locked P3 of the implicit group showed a significant region effect and harmony effect (all p values <.005). This indicates that the brain detected the rule violation at exactly 200ms, resulting in a P3 peak at 500ms (300ms after the violation). As for the explicit group, both stim-locked and resp-locked P3 showed a significant region effect ($p<.001$), but NOT a harmony effect (p values >.05). This means that the explicit group's detection of the rule violation was not reflected in P3. Furthermore, we found a significant lateralized readiness potential (LRP) in both groups, which reflects the response selection process.

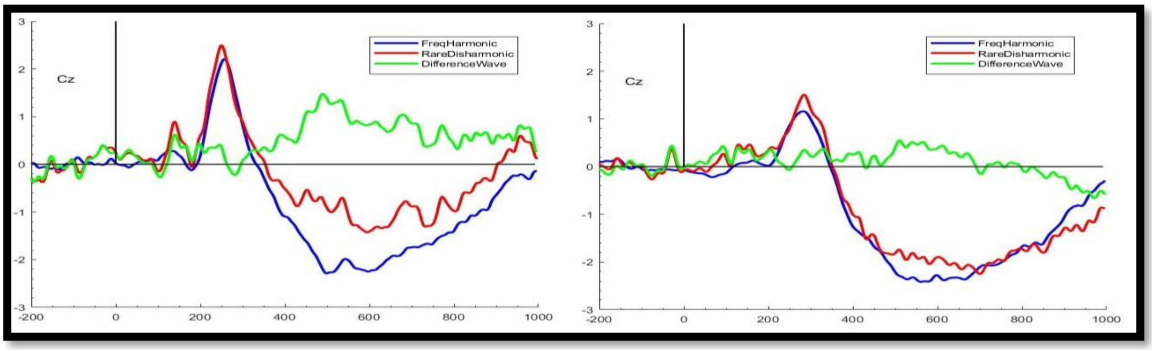

Fig. 1. Stimulus-locked grand average ERP waveforms: clear harmony effect reflected in P3 difference waveform in implicit (left panel) but not in explicit group (right panel). All stimuli elicited a clear auditory evoked potential (AEP).

Discussion and Conclusion. Behaviorally, both the explicit and implicit groups learned the non-adjacent phonotactic pattern, with the explicit group performing much better than the implicit group, reflected in both *d'* and accuracy. However, the two groups differed in their measured brain responses. The implicit learners showed a predicted P3 modulation to rule violation, while the explicit learners showed no modulation, despite the presence of a robust AEP and LRP in both groups. We interpret these results to indicate that implicit and explicit learning leads to different types of neural encoding of the acquired phonotactic rule. This interpretation is in line with Moreton et al. (2017)'s distinction between cue-based (implicit) and rule-based (explicit) models; the former is more like typical phonotactic learning whereas the latter is classic visual category learning which depends on frontal-striatal circuits (Ashby and Maddox, 2005) that may not be reflected on an EEG. Moreover, our results fit with the observation that first language acquisition is implicit and leads to long-term neural encoding, whereas adult 2^{nd} language acquisition, based on explicit learning, leads to a fundamentally weaker knowledge state. We conclude that lab-based learning experiments mimic naturalistic long-term implicit language learning.

Selected References. ♦ Ashby, F. G., & Maddox, W. T. (2005). Human category learning. Annu. Rev. Psychol., 56, 149-178. ♦ Domahs, Ulrike, W. Kehrein, J. Kraus, R. Wiese, and M. Schlesewsky. (2009). Event-related potentials reflecting the processing of phonological constraint violations. Language and Speech 52: 415–435. ♦ Finley, S. (2017). Locality and harmony: Perspectives from artificial grammar learning. Language and Linguistics Compass, 11(1), e12233. ♦ Luck, S. J., Kappenman, E. S., Fuller, R. L., Robinson, B., Summerfelt, A., & Gold, J. M. (2009). Impaired response selection in schizophrenia: evidence from the P3 wave and the lateralized readiness potential. Psychophysiology, 46(4), 776-786. ♦ Moore-Cantwell, Claire, J. Pater, R. Staubs, B. Zobel, and L. Sanders. (forthcoming). Event-related potential evidence of abstract phonological learning in the laboratory. ♦ Moreton, E., Pater, J., & Pertsova, K. (2017). Phonological concept learning. Cognitive science, 41(1), 4-69.