

Running head: SALIENCE EFFECTS IN INITIAL L2 ACQUISITION

**The effect of morphological salience and novel phonology in the initial stage of second language learning\***

Sara Fernández Santos<sup>1</sup>, Patrick Rebuschat<sup>2,3</sup>, Susana Correia<sup>4</sup>, Padraic Monaghan<sup>2</sup> & Miquel Llompart<sup>5,1</sup>

<sup>1</sup>Friedrich-Alexander University Erlangen-Nuremberg

<sup>2</sup>Lancaster University

<sup>3</sup>University of Tübingen

<sup>4</sup>NOVA University Lisbon

<sup>5</sup>Universitat Pompeu Fabra

**Address for correspondence:**

Friedrich-Alexander-Universität Erlangen-Nürnberg

Lehrstuhl für Language and Cognition (Alexander von Humboldt-Professur, Friedrich-Alexander University Erlangen-Nuremberg, Bismarckstr. 6, 91054 Erlangen.

[sara.fernandez@fau.de](mailto:sara.fernandez@fau.de)

**Acknowledgements:** We would like to thank Dr. Yuxin Ge and Dr. Yun-Wei Lee for their assistance with the experimental setup, and all members of the Lancaster Language Learning Lab for insightful discussions on statistical learning.

**Competing Interests:** The authors declare none.

## **The effect of morphological salience and novel phonology in the initial stage of second language learning**

### **Abstract**

This study investigates whether adult learners can simultaneously acquire the sounds, words, and grammar of a novel language through cross-situational statistical learning (CSL). English-speaking participants were exposed to an artificial language with unfamiliar phonology and varying morphological salience. Results showed measurable acquisition of nouns and verbs, but limited learning of adjectives and case markers. Importantly, only participants in the high-salience condition showed sensitivity to morphosyntactic violations, suggesting that perceptual salience enhances awareness of grammatical structure. A word–picture matching task revealed that learners encoded phonolexical forms imprecisely: while clearly deviant items were rejected, minimally different lures were often accepted. Finally, individual differences in phonolexical precision were only weakly associated with phonetic discrimination ability. These findings demonstrate the power and limits of CSL in second language learning and highlight the importance of perceptual cues for acquiring complex L2 structure.

*Keywords:* Cross-situational learning; implicit learning; statistical learning; perceptual salience; artificial language

In early second language acquisition (SLA), learners face the simultaneous challenges of segmenting unfamiliar phonological forms, mapping words to meaning, and extracting morphosyntactic regularities from continuous, variable input. Although phonological encoding, lexical learning, and grammar acquisition have been widely studied in isolation, much less is known about how they interact during initial exposure to a novel language. How do learners form stable form–meaning mappings when phonological forms are unfamiliar, lexical boundaries unclear, and grammatical cues opaque?

A central difficulty in early second language (L2) learning is referential ambiguity—the problem that words often co-occur with multiple potential referents (Quine, 1960). One mechanism for resolving this ambiguity is cross-situational learning (CSL), which enables learners to track statistical co-occurrences between linguistic forms and referents across multiple exposures. CSL has been shown to support the acquisition of various lexical categories across development, including nouns (Yu & Smith, 2007), verbs (Scott & Fischer, 2012), and combinations of both (Monaghan et al., 2015). More recently, it has been extended to grammar learning: Rebuschat et al. (2021) developed a CSL paradigm in which participants were exposed to an artificial language containing nouns, verbs, adjectives, and case markers. Participants viewed two dynamic visual scenes and selected the one that matched an auditory sentence in the artificial language. Crucially, no explicit feedback was provided, and participants were unaware that individual trials targeted different linguistic features (e.g., nouns, verbs, adjectives, case markers). This paradigm allowed researchers to test understanding of different linguistic features in a controlled, incidental learning environment. While lexical forms were learned relatively successfully, case marker acquisition remained at chance level, highlighting the difficulty of acquiring low-salience morphological features without explicit instruction.

In the present study, we adapt the paradigm introduced by Rebuschat et al. (2021) with two key innovations. First, we systematically manipulate the perceptual salience of case markers to test its impact on morphosyntactic learning. Second, we introduce L2-sounding pseudowords, thereby simulating early stages of L2 acquisition and increasing phonological variability. This approach reduces reliance on familiar lexical templates and aligns the learning task more closely with the demands faced by L2 learners in real-world contexts.

This design allows us to investigate whether CSL can support the joint acquisition of novel sounds, words, and grammatical structure in a typologically rich but under-researched language. While recent findings suggest that CSL can support L2 noun learning (Escudero et al., 2022; Ge, Monaghan, & Rebuschat, 2025a), this is, to our knowledge, the first study to systematically test whether cross-situational input is sufficient for acquiring an artificial language that integrates L2 phonology, lexical semantics, and morphosyntax.

### **The role of CSL in language learning**

CSL paradigms offer a powerful tool for studying how learners extract meaning from ambiguous input. Typically, participants are presented with two novel referents and one or more pseudowords, requiring them to track word–referent co-occurrences across trials to infer correct mappings. CSL thus operationalises statistical learning under referential ambiguity, a core challenge in naturalistic language learning.

In relation to SLA, CSL has been shown to support the acquisition of nouns containing L2 phonology, although accuracy tends to be lower when compared to learning in native-like phonotactic contexts (Escudero et al., 2022). This capacity remains across the lifespan, persisting into late adulthood (Ge et al., 2025b). Recent studies further demonstrate that CSL is robust enough to support the learning of words that differ in difficult L2 phonological contrasts (e.g., Escudero et al., 2022; Ge et al., 2025a), supporting its potential as a model for early stages of phonolexical acquisition. However, existing research has

focused almost exclusively on isolated word learning, leaving open the question of whether CSL can support more complex aspects of language learning, such as morphosyntactic development.

To address this, Rebuschat et al. (2021) developed a sentence-level CSL paradigm, described above, which captures a more naturalistic form of incidental learning. Studies using this paradigm have shown that learners can acquire multiple levels of linguistic structure simultaneously (Monaghan et al., 2021; Walker et al., 2020), with learning outcomes that mirror developmental sequences in natural language acquisition—most notably, the earlier acquisition of content words relative to function morphemes (Rebuschat et al., 2021). Furthermore, these effects are not transient: learning persists over time and correlates with memory systems known to underlie naturalistic SLA (Walker et al., 2020).

A major challenge in L2 acquisition, particularly in immersive contexts, is unfamiliar phonology. While novel form representations may be encoded rapidly, their precision often declines when L2 phonemes diverge from the learner's L1 inventory. This can lead to underspecified or fuzzy representations (Bordag et al., 2022), particularly when multiple L2 sounds are assimilated to a single L1 category—for example, English /æ/ and /ɛ/ both mapping onto German /ɛ/ (Llompart, 2021). Such fuzziness disrupts the mapping between phonological and semantic representations and introduces processing delays, comprehension errors, and reduced learning efficiency.

Against this backdrop, the first aim of the current study is to investigate whether CSL can support the simultaneous acquisition of L2 words and grammar from continuous sentence input, using a phonology that differs typologically from participants' L1. By incorporating L2-like pseudowords and longer speech streams, we seek to approximate early L2 acquisition in ecologically valid learning conditions—especially those that characterise immersion environments, where learners face high referential ambiguity and minimal explicit

instruction. Additionally, we examine the phonological granularity of the resulting lexical representations, testing whether learners encode fine phonemic detail in their newly formed L2 words and how this relates to semantic accuracy. This approach allows us to bridge research on CSL, phonolexical encoding, and morphosyntactic development within a unified experimental framework.

### **The role of salience in the acquisition of L2 morphology**

Morphological case marking remains one of the most persistent challenges in adult language learning. Even intermediate-level learners often struggle to use case consistently (Haznedar, 2008), a pattern replicated in artificial language learning studies (e.g., Fedzechkina et al., 2017; Kenanidis et al., 2023; Rebuschat et al., 2021). Several factors contribute to the difficulty of acquiring L2 morphology, including low perceptual salience, redundancy of form–function mappings, lack of transparency, and crosslinguistic transfer (Ellis, 2017).

In this study, we focus on perceptual salience, defined as an intrinsic property of the linguistic form itself—distinct from contextual, semantic, or learner-related notions of salience (Boswijk & Coler, 2020; Knell et al., 2025). Goldschneider & DeKeyser (2001) propose that, in the domain of morphology, perceptual salience can be operationalised via segmental length, syllabicity, sonority, stress, and positional prominence. Their meta-analysis of English L2 morphology acquisition showed that perceptual salience was the strongest predictor of acquisition order ( $r = .63$ ), underscoring its foundational role.

Theoretical models of SLA have long emphasised the role of attention in form learning. According to the Noticing Hypothesis (Schmidt, 1990), learners must consciously attend to linguistic forms for these to be integrated into developing interlanguage systems. Increased perceptual salience is thought to enhance noticing by making relevant cues more available to awareness (Williams & Rebuschat, 2023). Supporting this view, Simoens et al.

(2017) manipulated case marker salience in a semi-artificial language during meaning-focused versus grammar-focused reading tasks. Highly salient markers were skipped less often and elicited longer fixations in ungrammatical contexts—especially under explicit learning conditions—indicating greater attention allocation and deeper processing. These effects were not observed for low-salience markers, suggesting that form characteristics themselves modulate perceptual uptake.

In a recent paper, Kenanidis et al. (2024) provided evidence from a fully artificial language learning paradigm which further suggests that increasing the perceptual salience of morphological markers can promote learning even in the absence of explicit instruction (Kenanidis et al., 2024). However, this study conflates salience with marker number—the high-salience condition used a single agentive marker (*-pazz*), whereas the low-salience condition featured two markers (*-i* and *-o*)—thereby limiting the interpretability of its findings.

The second aim of the present study was therefore to isolate and test the effect of perceptual salience on L2 morphological acquisition under tightly controlled conditions. Using an artificial language learning paradigm, we systematically manipulated the salience of two morphological case markers (agent and patient) across four phonological dimensions: length, syllabicity, stress and sonority. Low-salience markers (*/-l/*, */-r/*) were monosegmental, non-syllabic, unstressed, and contained liquid consonants, which are closest to vowels in the sonority scale except for glides (Clements, 1990; Selkirk, 1984)<sup>1</sup>. In contrast, high-salience markers (*/-'ka/*, */-'ji/*) were multisegmental, syllabic, stressed, and composed of phonologically distinct consonant–vowel sequences including obstruents, which are substantially more distant from the preceding vowel on the sonority scale. This design allows

---

<sup>1</sup> Liquid consonants are articulated with a weaker constriction and, therefore, a relatively unobstructed airflow compared to other consonants and closer to vowels in that regard. This, together with their some of their combinatoric properties with regard to syllable structure (e.g., */l/* and */r/* can be syllabic in some languages, like Czech and Slovak) has led to them being situated right after vowels and glides in the sonority hierarchy.

us to test whether increasing perceptual salience enhances incidental morphological learning without conflating it with marker frequency, function, or form complexity.

By embedding this manipulation within a meaning-focused cross-situational learning task, we directly assess whether perceptual salience alone can facilitate the acquisition of otherwise challenging grammatical forms in an L2 learning context.

### **The present study**

This study's design, hypotheses, and analysis plan were preregistered ([anonymised OSF link](#)).

We exposed adult native English speakers to an artificial language based on the phonotactics and morphology of Portuguese—a global language that remains underrepresented in SLA research (Plonsky, 2023). The artificial language included nouns, verbs, adjectives, and morphological case markers that varied in perceptual salience (see previous section), and the sentences had a verb-final structure. Learners were exposed to the language through a CSL paradigm in which they mapped spoken sentences to animated scenes across multiple exposures. Comprehension of vocabulary was tested within the CSL task itself, while morphosyntactic awareness was assessed using a grammaticality judgement task. To evaluate the phonological specificity of lexical representations, we included a word–picture matching (WPM) task featuring phonological lures, followed by an AX phoneme discrimination task aimed at 4 phonological contrasts of Portuguese that are known to be challenging for L1-English learners.

Our pre-registered hypotheses focus on the effect of salience on L2 morphology acquisition. Particularly, we hypothesised that participants exposed to high-salience case markers would outperform those in the low-salience condition in their comprehension of case morphology (H1). In addition to improved grammatical performance, we expected salience to yield knock-on effects on lexical learning. Specifically, we hypothesised that the high-

salience group would show higher accuracy scores on overall learning through the blocks (H2), more accurate acquisition of nouns (due to co-occurrence with salient markers; H3), verbs (due to enhanced understanding of argument structure; H4). Similarly, we hypothesised that the high-salience group would show better detection of word order violations (due to clearer segmentation of morphemes; H5), and of case marker violations (because the greater salience should make the structure more prominent; H6). The potential effect of salience condition on adjective learning was explored without a directional hypothesis.

Additionally, we pre-registered as exploratory analyses those that correspond to the WPM task and AX task, which measured the phonetic detail of the acquired representations and their relation to phoneme discrimination ability. Particularly, we investigated whether the lexical representations formed during CSL are phonologically specific or fuzzy with the WPM task. In this regard, we expected that while learners would reliably reject pseudowords with phonological substitutions that are substantially different from the learned forms, they would struggle to reject minimally different lures where the target sounds substituted are expected to be confusable, suggesting that the phonolexical representations formed during CSL remain partially underspecified. We also asked whether individual differences in phonolexical precision are linked to L2 phoneme discrimination abilities. For this question, we expected that learners who were better able to discriminate challenging L2 phonological contrasts in the AX task would also show greater accuracy in rejecting 1-feature mismatches in the WPM task (which corresponded to the same contrasts tested in the AX task), indicating that more precise phonetic sensitivity supports more accurate phonolexical encoding.

Regarding the question of whether adults can acquire novel words and grammar simultaneously by tracking cross-situational statistics even when the language has a non-native phonology, we expected participants to successfully learn both lexical items

(particularly nouns and verbs) and grammatical structure (e.g., case assignment and word order), despite the L2 phonological properties of the input.

## Methods

### Participants

One hundred and sixty participants ( $M_{age} = 38.2$ ;  $SD_{age} = 11.4$ ; female = 83) took part in the study in exchange for financial compensation. Half of the participants ( $n = 80$ ) were assigned to the high salience condition and the remaining half ( $n = 80$ ) to the low salience condition. All participants were self-identified English monolinguals, and none reported knowledge of Portuguese nor any verb-final language. Participants were recruited using the online platform Prolific ([www.prolific.co](http://www.prolific.co)). The study was approved by the ethics review panel of the Faculty of Arts and Social Sciences of Lancaster University and conducted in accordance with the Declaration of Helsinki. All participants provided informed consent before their participation.

Sample size was calculated through a power analysis consisting of running simulations of the planned mixed effects model for the main effect of salience, which, based on previous studies, was estimated to produce a difference of around 10% in case marker accuracy by the end of the study. This estimation was based on data from two artificial language experiments. From Kenanidis et al. (2024), we compared accuracy in the last block of grammar test trials between a group with only one salient patient marker (*-pazz*, 76.44%) and one with two less salient patient and agent case markers (*-i* and *-o*, 62.67%). From Zhu et al. (2025), we compared accuracy in the last block of number marker test trials between three groups differing in the salience of number markers (high salience (plural and singular markers): 60%; low salience (plural and singular markers): 43%; low salience (plural) and no

marker (singular): 50%). After running 100 simulations, the power estimation surpassed .80 for the main effects of salience with 80 participants per condition, i.e. 160 total.

### **Materials: Artificial language**

Participants were trained and tested on an artificial language designed to incorporate novel phonological, lexical, and grammatical features. The language consisted of 14 disyllabic pseudowords: 8 nouns (each referring to a unique alien character), 4 verbs (representing transitive actions), and 2 adjectives (corresponding to alien colours). In addition, two obligatory case markers were implemented as suffixes on nouns to indicate grammatical roles (agent and patient). All pseudowords adhered to Portuguese phonotactics and followed a CVCV structure, with either the first vowel or the second consonant corresponding to one of four L2 phonemes: /k/, /ɲ/, /e/, or /o/.

The perceptual salience of case markers was systematically manipulated (high-salience: /-ʃi/ and /-ka/, low-salience: /-l/ and /-r/). Details of this manipulation are provided in the introduction. To minimise potential sound–meaning biases, four different pseudoword-to-referent mappings were created and randomly assigned to participants. The same lexicon was used consistently across all learning and testing phases. All stimuli are available in the supplementary materials.

The pseudowords were recorded in isolation by a female native speaker of European Portuguese. Noun + suffix combinations were recorded as single lexical items. Recordings were made using a Tascam DR-22WL digital recorder and a Monacor HSE-130/SK head-mounted microphone at 44.1 kHz, 16-bit resolution. A 250 ms silence was appended to the end of each file using Audacity® (Audacity Team, 2024). Sentences were constructed through concatenation<sup>2</sup> of these pseudowords in Gorilla Game Builder.

---

<sup>2</sup> The concatenation of pseudowords with fixed silences may result in quasi-metronomic streams that enhance the perceptual prominence of each lexical item and, potentially, of their suffixes (i.e., case markers). However, note that this applies to the low-salient and high-salient items alike. For this reason, we believe it is unlikely that this has a major effect on any between-group comparisons.

The artificial language used two sentence structures: SOV and OSV. Optional pre-nominal adjectives occurred in half of the noun phrases and were counterbalanced across subject and object positions. Both nouns in each sentence were always marked for case. Sentence structures followed those of Rebuschat et al. (2021), with vocabulary testing expanded to include more trials. In total, 192 training (exposure) sentences and 128 vocabulary test sentences were constructed. Eight fully counterbalanced stimulus lists were created by crossing the four pseudoword–referent mappings with the two salience conditions.

Example sentences for the high- and low-salience conditions are shown in (1) and (2) respectively:

(1) / (tiɫu) meguʻʃi (piɲu) kiɫuʻka suɫu /

(2) / (tiɫu) megul (piɲu) kiɫur suɫu /

(red) megu-ACC (blue) kiɫu-NOM jump-over

‘The (blue) kiɫu is jumping over the (red) megu.’

For the grammaticality judgement task, an additional 48 sentences were constructed: 24 grammatical and 24 ungrammatical. Ungrammatical items contained either incorrect verb placement (as in Rebuschat et al., 2021) or incorrect case marker assignment (newly introduced). See Table 1 for a summary of grammatical and ungrammatical structures.

The visual referents used for the CSL task were identical to those developed by Rebuschat et al. (2021). They included eight distinct cartoon-style alien characters, each rendered in either red or blue, and animated using Gorilla Game Builder to perform one of four transitive actions: hiding behind, jumping over, pushing, or lifting another character.

**Table 1**

*Grammatical and ungrammatical sentence types used in the grammaticality judgement task*

Sentence Type	Violation Type (if applicable)	Syntactic Pattern	Nr of Trials
---------------	--------------------------------	-------------------	--------------

---

Grammatical	n/a	NP <sub>NOM</sub> NP <sub>ACC</sub> VP	12
	n/a	NP <sub>ACC</sub> NP <sub>NOM</sub> VP	12
Ungrammatical	Verb placement error	VP NP <sub>ACC</sub> NP <sub>NOM</sub>	2
		VP NP <sub>NOM</sub> NP <sub>ACC</sub>	2
		NP <sub>NOM</sub> VP NP <sub>ACC</sub>	2
		NP <sub>ACC</sub> VP NP <sub>NOM</sub>	2
	Case misassignment	NP <sub>ACC</sub> NP <sub>ACC</sub> VP	2
		NP <sub>NOM</sub> NP <sub>NOM</sub> VP	2
	Case marker omission	NP (-) NP <sub>ACC</sub> VP	2
		NP <sub>ACC</sub> NP (-) VP	2
		NP <sub>NOM</sub> NP (-) VP	2
		NP (-) NP <sub>NOM</sub> VP	2

---

### Tasks and procedure

Participants completed the study online via Gorilla ([www.gorilla.sc](http://www.gorilla.sc)). After providing informed consent, they filled out the revised version of the Language and Social Background Questionnaire (LSBQ; Anderson et al., 2018) to report their linguistic and educational background. The experimental session consisted of four tasks, presented in fixed order: (1) the cross-situational learning (CSL) task, (2) the grammaticality judgement task (GJT), (3) the word–picture matching (WPM) task, and (4) the AX phoneme discrimination task. Finally, participants provided a brief retrospective verbal report to assess their awareness of the morphological case markers. The full experiment lasted approximately 70 minutes. All materials and data are available via the [anonymised OSF repository](#).

#### *Cross-Situational Learning Task*

Each CSL trial presented participants with two animated scenes side by side, accompanied by a spoken sentence in the artificial language. Participants were instructed to select the scene that best matched the sentence by pressing ‘Q’ (left scene) or ‘P’ (right scene) on their keyboard. The only instruction given was: “Choose the scene that matches the sentence.”

The CSL task consisted of two trial types: exposure trials, designed to promote learning, and test trials, designed to assess learning.

**Exposure trials.** In exposure trials, the two scenes differed in two or more aspects (e.g., both the verb and the identity of the agent). This preserved referential ambiguity while enabling participants to track statistical regularities across trials. For example, one scene might show a red alien pushing a blue alien, while the other shows a blue alien lifting a red alien. In this way, multiple cues (e.g., verb, agent, patient, colour) co-varied, allowing participants to incrementally associate spoken forms with their meanings through cross-situational inference. See Figure S1 in the supplementary materials for an example.

**Test trials.** Test trials were identical in structure and appearance to exposure trials but served to probe what participants had learned. The critical difference was that in test trials, the two scenes differed in only one specific aspect, creating a minimal pair. This manipulation allowed us to isolate participants’ knowledge of particular lexical or grammatical elements. No indication was given that testing was taking place.

There were four types of test trials. In noun trials, the two scenes differed only in the identity of one alien character (subject or object). In verb trials, the scenes depicted the same characters but performing different actions. In adjective trials, one alien differed in colour (red vs. blue). Finally, in marker trials, the agent–patient roles were reversed, keeping all other elements constant. For example, in a marker test trial, one scene might show a red alien

pushing a blue alien, and the other a blue alien pushing a red alien—differing only in thematic role assignment. See Figure S2 in the supplementary materials for examples.

**Structure.** The CSL task comprised 8 exposure blocks and 4 mixed blocks, with each mixed block including both exposure and test trials. This structure enabled us to monitor the trajectory of learning while maintaining uniform task presentation.

Each exposure block contained 16 exposure trials. Mixed blocks added a further 16 exposure trials and 32 test trials each, yielding a total of 192 exposure trials and 128 test trials across the task. Among the exposure trials, 170 sentences were unique and 11 were repeated once (never consecutively).

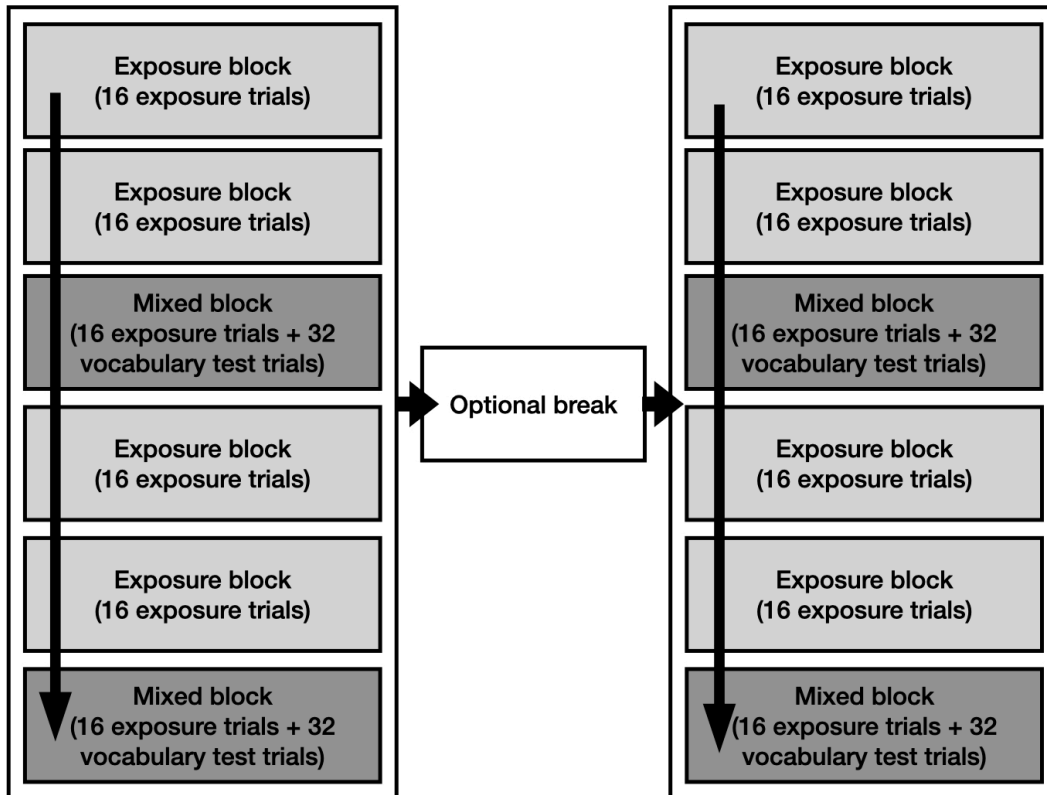
Test trials were distributed as follows per mixed block: 16 marker trials, 8 noun trials, 4 verb trials, and 4 adjective trials. Marker test trials included 32 unique sentences, each presented twice (non-consecutively) within a block, resulting in 64 marker items overall. The remaining test trials (noun, verb, adjective) comprised 56 unique sentences and 4 repeated ones. None of the test sentences had occurred during exposure.

This design allowed for a fine-grained assessment of vocabulary and morphosyntactic learning while keeping the format and appearance of test trials identical to exposure trials. Figure 1 provides a visual overview of the CSL structure.

Across all blocks, sentence presentation was balanced for word order, the presence of adjectives, noun and verb frequency, and agent–patient roles. Participants were not informed of any properties of the language and received no explicit instruction or feedback at any point.

**Figure 1**

*Structure of the cross-situational learning (CSL) task. The task consisted of 8 exposure blocks and 4 mixed blocks combining exposure and test trials. Each exposure block included 16 exposure trials. Each mixed block included 16 exposure trials and 32 test trials targeting vocabulary and case marker learning.*



### ***Grammaticality judgement task***

We used a grammaticality judgement task (GJT) to assess participants' knowledge of morphosyntactic patterns after the CSL task. In each trial, participants viewed a single animated scene while listening to a spoken sentence and judged whether the sentence was grammatical or ungrammatical. Responses were made by pressing 'Q' ("weird") or 'P' ("good") on the keyboard. See Figure S3 in the supplementary materials for an example trial. The task included 48 trials, presented in randomised order. Half of the sentences were grammatical and counterbalanced for word order; the remaining half were ungrammatical, featuring violations of case marking or word order (see Table 1). None of the sentences had been presented during the CSL task.

### ***Word-Picture Matching (WPM) Task***

The WPM task was used to assess the phonetic specificity of lexical representations formed during the CSL task. In particular, it tested whether participants encoded fine-grained phonological detail for the L2 phonemes /k/, /ɲ/, /e/, and /o/, or whether their representations were fuzzy and underspecified. It also provided an additional measure of form–meaning mappings acquired during learning.

During the task, participants saw an isolated visual referent (noun, adjective, or verb) from the CSL task for 2 seconds, followed by a spoken pseudoword. They were instructed to judge whether the spoken word matched the referent learned during CSL. Responses were recorded via keypress: 'Q' for "incorrect word" and 'P' for "correct word."

The spoken stimuli comprised four conditions designed to assess the phonological specificity of lexical representations. In the *match* condition, the pseudoword was the correct label for the visual referent learned during the CSL task (e.g., /kiʎu/ for the alien). The *one-feature mismatch* (1FM) condition presented a pseudoword that differed from the target by a single phonological feature, crucially forming one out of 4 challenging L2 contrasts: /l/–/ʎ/,

/n/–/ɲ/, /e/–/ɛ/, and /o/–/ɔ/ (e.g., /kɪlu/). In the *three-feature mismatch* (3FM) condition, the pseudoword differed from the target by three phonological features (e.g., /kɪfu/). Particularly, for consonant contrasts, the 1FM pseudowords differed from the target only in place of articulation, whereas the 3FM pseudowords differed in place of articulation, manner of articulation and voicing; and for the vowel contrasts, the 1FM pseudowords differed only in vowel height, whereas the 3FM pseudowords differed in vowel height, tongue position and rounding (see Supplementary Materials, Table S3). Finally, the *control* condition used a pseudoword that corresponded to a completely different referent, unrelated to the one presented (e.g., /dɪɲu/).

To ensure participants were familiar with the referents in isolation (as they had only encountered them in sentence context during CSL), three introductory screens presented all referents by category: alien characters (black and white line drawings), colour squares, and animated actions. Each screen was accompanied by a reminder (e.g., “During this experiment, you have learned the words for these [aliens/colours/actions]”).

The task comprised 84 trials presented in random order. Each correct mapping was presented three times to balance the number of trials across conditions. An example of the visual layout is shown in Figure S4 in the supplementary materials.

### ***AX Discrimination Task***

Given the similarities between the L2 sounds in the match and 1-feature mismatch conditions of the WPM task (e.g., [kɪʎu]–[kɪlu]), we expected low accuracy in detecting the mismatch when accessing the newly acquired representations. The purpose of the AX discrimination task was to determine whether participants could perceptually distinguish between these challenging L2 contrasts—and thereby assess the extent to which low accuracy in the WPM task stemmed from difficulty in perceiving these phonological differences.

The task, adapted from Ge et al. (2024), involved 12 disyllabic pseudoword minimal pairs, with three items per contrast: /l/–/ʎ/, /n/–/ɲ/, /e/–/ɛ/, and /o/–/ɔ/. Pseudowords were sourced from Correia et al. (2025). They were recorded by three native Portuguese speakers—two female and one male—including the speaker who had recorded the CSL stimuli. Most pseudowords were novel and had not appeared in the previous tasks, with the exception of three items (two from the /n/–/ɲ/ contrast and one from the /o/–/ɔ/ contrast). In each pair, the two pseudowords were always produced by different speakers, and each minimal pair was recorded twice, using different speaker combinations.

During the task, participants listened to a pair of pseudowords and judged whether the two items sounded the same or different by clicking on “SAME” or “DIFFERENT.” The task comprised 48 trials presented in randomised order. Half of the trials contained identical pairs (e.g., [kiʎu]–[kiʎu] or [kilu]–[kilu]), and half contained different pairs (e.g., [kiʎu]–[kilu] or [kilu]–[kiʎu]).

### ***Retrospective verbal reports***

At the end of the experiment, participants completed a short retrospective questionnaire to assess their awareness of the case-marking system in the artificial language (Rebuschat, 2013). They were asked open-ended questions about the strategies they used during the CSL task, whether their approach changed over time, and whether they noticed any patterns or regularities in the language, such as recurring word endings (e.g., “-L” and “-R”). To quantify awareness, we calculated an awareness score for each participant. One point was awarded for each marker correctly identified as a suffix during the first half of the CSL task, and 0.5 points for markers identified during the second half, yielding a maximum possible score of 2.

## **Analysis**

## Pre-registered analyses

All analyses were conducted in R. For all the mixed-effects models, the random-effects structure was defined as maximally as possible without causing convergence issues (see supplementary materials for full model specifications).

To explore H1, which stated that the high-salience condition would outperform the low-salience condition in case marker learning, we conducted a generalised linear mixed-effects model with a logit link function (using the lme4 package; Bates et al., 2015). The dependent variable was accuracy on case marker test trials in the final block of the CSL task (1 = correct, 0 = incorrect). The model included salience condition as a fixed effect, with random intercepts for participants and items. We refer to this model henceforth as GLMER1.

To explore H2, which stated that perceptual salience of case markers would influence overall learning across the CSL task, we conducted a generalised linear mixed-effects model with accuracy on all exposure trials as the binary dependent variable (1 = correct, 0 = incorrect). Fixed effects included salience condition, block (1–12), and their interaction. The model included random intercepts for participants and items, and random slopes for block over participants. We refer to this model henceforth as GLMER2.

To test H3 and H4, which stated that the high salience markers would positively impact the acquisition of nouns and verbs respectively, and to explore the possible knock-on effect on adjective learning, we ran separate models for each category using test trial accuracy from the final CSL block. All models included salience condition as a fixed effect and random intercepts for items and participants. We refer to these models as GLMER3 (nouns), GLMER4 (verbs) and GLMER5 (adjectives).

To test the hypotheses that case marker salience would benefit word order acquisition (H5) and marker syntax (H6), we computed d-prime ( $d'$ ) scores on the GJT for each

participant, separately for word order and case marker trials respectively. We then ran two linear regression analyses to test the effect of salience on these  $d'$  scores. We refer to these models as LM1 (word order) and LM2 (case markers).

As pre-registered exploratory analyses, we first investigate the nature of the acquired lexical representations. For this, we ran a mixed-effects logistic regression on trial-level data of the WPM task. The dependent variable was accuracy. Fixed effects included mismatch condition, salience condition, target phoneme, referent type (Match pseudoword's function during the CSL task) and the interaction between mismatch condition and target phoneme. Random intercepts were included for participant and item, with random slopes for salience over item and referent type over participant. We refer to this model as GLMER6. We also explore whether individual differences in phonetic discrimination ability help explain the phonological fuzziness observed in word learning. With this purpose, we ran another mixed-effects logistic regression on data from the AX task and WPM task trials on 1-feature mismatches. The dependent variable was accuracy. Fixed effects included task, target phoneme, their interaction, and salience condition. Random intercepts were included for participant and item, with random slopes for item over participant. We refer to this model as GLMER7.

Variable coding was pre-registered, and from this we kept the coding for variables "salience condition" (high = 0.5; low = -0.5) and "task" (word-picture matching task = 0.5; AX discrimination task = -0.5). Due to model convergence issues, the coding scheme was changed and centred for variables "target phoneme" (sum-coding) and "mismatch condition" (custom coding via hypothesis matrix: the first contrast compares the Match condition with Control, the second contrast compares Match with 1FM and the third contrast compared Match with 3FM). The variable of "referent type" in GLMER6, added upon reviewer request,

was sum-coded. All the variables introduced in the additional analyses reported below were sum-coded.

### **Outlier removal**

We followed our pre-registered plan for data removal. For all tasks, we removed individual reaction times over 15s. In the CSL task, we removed data from blocks where a participant produced the same response (e.g., left side key) or alternating pattern (e.g. left/right/left/right) for 80% or more trials. This removed 0.03% of the total data.

### **Additional analyses**

To explore the question of whether adults can simultaneously acquire a novel grammar and lexicon via CSL, we examined whether predicted accuracy differed from chance (50%) in each condition for exposure trials, vocabulary test trials and grammaticality judgements. For exposure trials, we computed estimated marginal means (EMMs) from GLMER2 using the emmeans package in R. EMMs were extracted on the response (probability) scale, and Wald z-tests were conducted against a null of 0.5 (chance level). P-values were adjusted for multiple comparisons across blocks using the Holm procedure for these and all the following z-tests. For vocabulary test trials, we similarly computed EMMs from GLMER1 (case markers), GLMER3 (nouns), GLMER4 (verbs), and GLMER5 (adjectives) and conducted Wald z-tests against a null of 0.5. Lastly, for grammaticality judgements we computed EMMs on LM1 (word order - GJT) and LM2 (case marker - GJT) and conducted one-sample t-tests on  $d'$  scores against a null of 0, reflecting chance level of  $d'$ .

Upon reviewer request, we included an additional linear model to compare self-reported awareness across the two salience conditions. As a supplementary signal-detection measure in the WPM task, we also computed  $d'$  scores to probe discrimination between Match and 1FM pseudowords. We further conducted a series of Bonferroni-adjusted

correlations between self-reported awareness and LSBQ scores with vocabulary performance and  $d'$  scores in the GJT to explore whether retrospective awareness of the case marking suffixes or language background were related to vocabulary or grammar learning.

Additionally, we examined via a Pearson correlation whether phoneme discrimination ability (AX task) predicted performance in the WPM task, specifically the ability to reject 1-feature mismatches, a key indicator of phonolexical specificity.

## Results

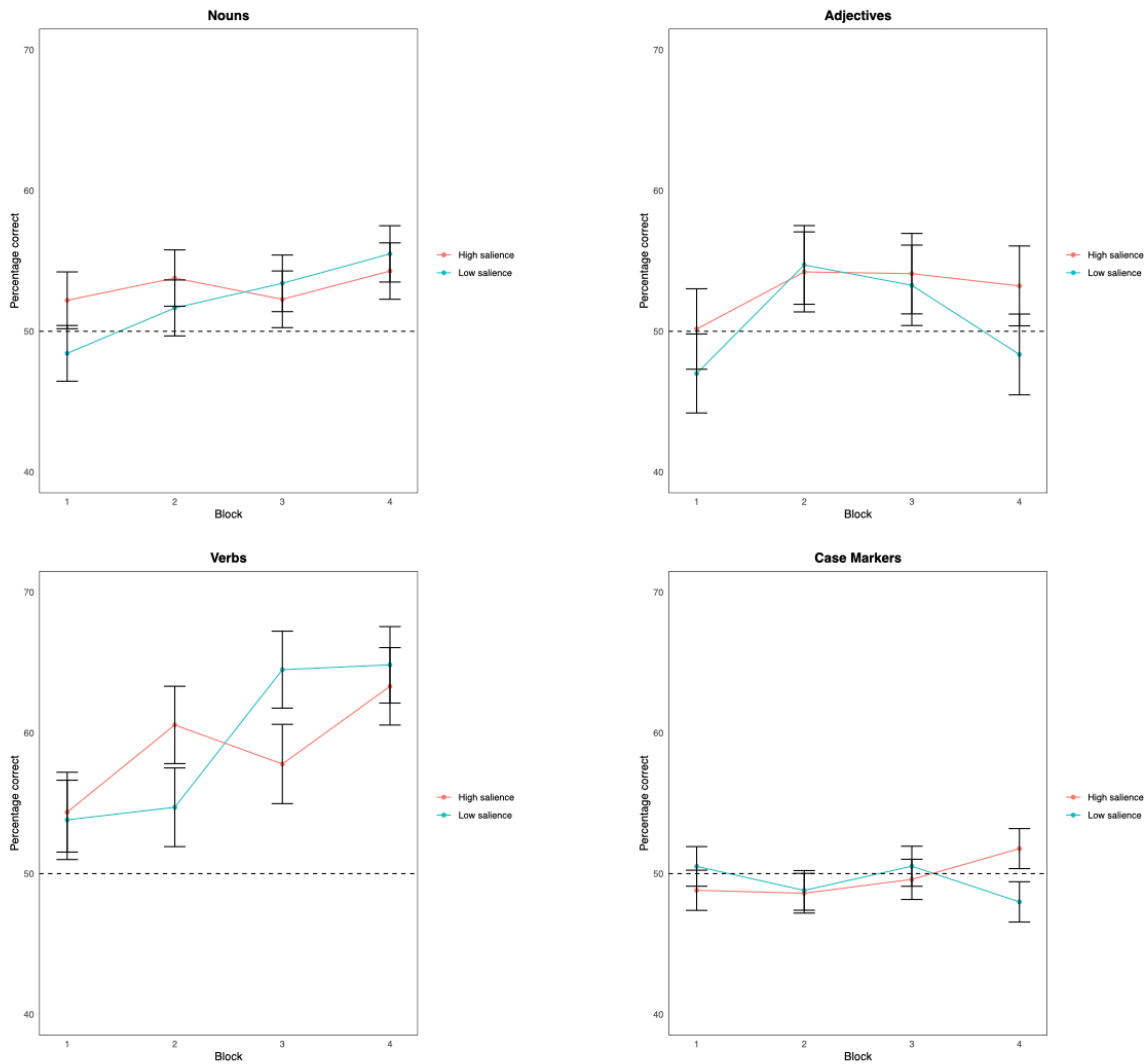
All the statistical tests reported in this study were run with R studio (R Core Team, 2017; R version 4.2.2). The data and code scripts for the analyses reported in this study can be accessed on this [anonymised OSF site](#).

### Cross-Situational Learning Task

The detailed performance on the exposure trials of the CSL task is reported in the supplementary materials (Figure S5). Accuracy for the four types of test trials is shown in Figure 2.

### Figure 2

*Percentage of correct responses across four test blocks in vocabulary tests for nouns, adjectives, verbs, and case markers, by salience condition (high vs. low). The graphs are scaled from 40% to 70% accuracy for clearer visualisation. The dashed horizontal line indicates chance performance (50%). Error bars represent standard errors of the mean.*



GLMER1 was run to test H1, which stated that the high-salience group would perform more accurately than the low-salience group in case marker vocabulary test trials in the last block. The results showed a marginal effect of salience in the predicted direction: participants in the high-salience condition were more accurate than those in the low-salience group ( $b = 0.15$ ,  $SE = 0.08$ ,  $z = 1.83$ ,  $p = .07$ ), although the effect did not reach conventional significance. Model fit statistics indicated that the fixed effect of salience explained very little of the variance (marginal pseudo- $R^2 < .01$ ; conditional pseudo- $R^2 < .01$ ). To explore whether

performance was above chance at the last block of case marker vocabulary test, we computed EMMs on GLMER1 for each condition and compared these against chance (0.5). The results showed that model-based predicted accuracy in the final block did not differ significantly from chance in either the low-salience condition ( $M = .48$ ,  $SE = .02$ ,  $z = -1.31$ ,  $p = .19$ ) or the high-salience condition ( $M = .52$ ,  $SE = .02$ ,  $z = 1.17$ ,  $p = .24$ ).

GLMER2 was run to test H2, which stated that the high-salience group would perform more accurately than the low-salience group on training trials across the CSL task. The results revealed a significant effect of block ( $b = 0.10$ ,  $SE = 0.01$ ,  $z = 7.29$ ,  $p < .001$ ), indicating that learning improved over time. While the effect of salience did not reach statistical significance ( $b = 0.09$ ,  $SE = 0.06$ ,  $z = 1.51$ ,  $p = .13$ ), the trend suggested that participants in the high-salience condition may have outperformed those in the low-salience condition. The fixed effects explained 2% of the variance (marginal pseudo- $R^2 = .02$ ), with total variance explained at 25% (conditional pseudo- $R^2 = .25$ ).

To test overall learning in the CSL task, we extracted EMMs on GLMER2 for each condition across blocks and for each condition at each block. EMMs were compared against chance (0.5) using Wald z-tests. Across all blocks, predicted accuracy was significantly above chance for both conditions (low salience:  $M = .63$ ,  $SE = .03$ ,  $z = 4.86$ ,  $p < .001$ ; high salience:  $M = .64$ ,  $SE = .03$ ,  $z = 5.27$ ,  $p < .001$ ). When examining block-wise performance, predicted accuracy was near chance in the initial blocks and increased over time, with both conditions exceeding chance at later blocks (high salience: from block 2 onwards, low salience: from block 3 onwards). Holm-adjusted p-values were used to correct for multiple comparisons across the twelve blocks. See supplementary materials, Table S5 for the results of the Wald z-test comparisons for each condition at each block.

To test whether salience affected the acquisition of nouns and verbs, as predicted by H3 and H4 respectively, as well as to explore the potential impact of salience group on

adjective learning, we ran separate models for each category using test trial accuracy from the final CSL block: GLMER3 (nouns), GLMER4 (verbs) and GLMER5 (adjectives).

None of the models yielded significant effects. For noun learning, there was no effect of salience and a slight trend opposite to predictions ( $b = -0.05$ ,  $SE = 0.11$ ,  $z = -0.43$ ,  $p = .66$ ; marginal  $R^2 < .01$ ). The same was true for verbs ( $b = -0.07$ ,  $SE = 0.24$ ,  $z = -0.31$ ,  $p = .75$ ; marginal  $R^2 < .01$ ), despite moderate variance explained at the participant level (conditional  $R^2 = .17$ ). For adjectives, the salience effect also fell short of significance ( $b = 0.20$ ,  $SE = 0.16$ ,  $z = 1.21$ ,  $p = .23$ ), though the trend was in the expected direction. Wald z-tests against chance run on the EMMs of GLMER3, GLMER4 and GLMER5 showed that predicted accuracy for both salience groups was above chance in the last vocabulary test block for nouns (low salience:  $M = .55$ ,  $SE = .02$ ,  $z = 2.66$ ,  $p < .01$ ; high salience:  $M = .54$ ,  $SE = .02$ ,  $z = 2.07$ ,  $p < .05$ ) and verbs (low salience:  $M = .67$ ,  $SE = .04$ ,  $z = 4.29$ ,  $p < .001$ ; high salience:  $M = .66$ ,  $SE = .04$ ,  $z = 3.87$ ,  $p < .001$ ). In contrast, predicted accuracy for both salience groups in the last vocabulary test block for adjectives was not significantly different from chance (low salience:  $M = .48$ ,  $SE = .03$ ,  $z = -0.47$ ,  $p = 0.64$ ; high salience:  $M = .53$ ,  $SE = .03$ ,  $z = 0.92$ ,  $p = .35$ ).

Although not part of our preregistered analyses, we explored whether retrospective awareness of the case marking suffixes was related to learning outcomes. Particularly, we run correlation tests between awareness scores and LSBQ scores each with both overall vocabulary and case marker vocabulary performance. We report Bonferroni-corrected p-values. Awareness scores, based on participants' verbal reports, were not significantly correlated with overall vocabulary performance in the final block after Bonferroni corrections, but the data shows a trend in the expected direction ( $r(154) = .19$ ,  $p = .06$ ). However, awareness scores correlated with performance on case marker trials only ( $r(154) = .24$ ,  $p < .05$ ). In contrast, composite scores from the LSBQ were not significantly correlated

with either overall vocabulary performance ( $r(154) = -.09, p = .94$ ) or case marker accuracy ( $r(154) = .04, p = 1$ ).

Additionally, we run a linear model to test whether awareness scores differed significantly between the two salience groups. This model revealed a significant effect of salience condition on awareness scores, with the low-salience group showing lower awareness scores than the high-salience group ( $b = -0.27, SE = 0.01, t = -47.57, p < .001$ ). Cohen's  $d$  reveals that this is a moderate effect size ( $d = .39$ ), although the model indicates that approximately only 3.8% of the variance in awareness is explained by salience group ( $R^2 = .038$ ).

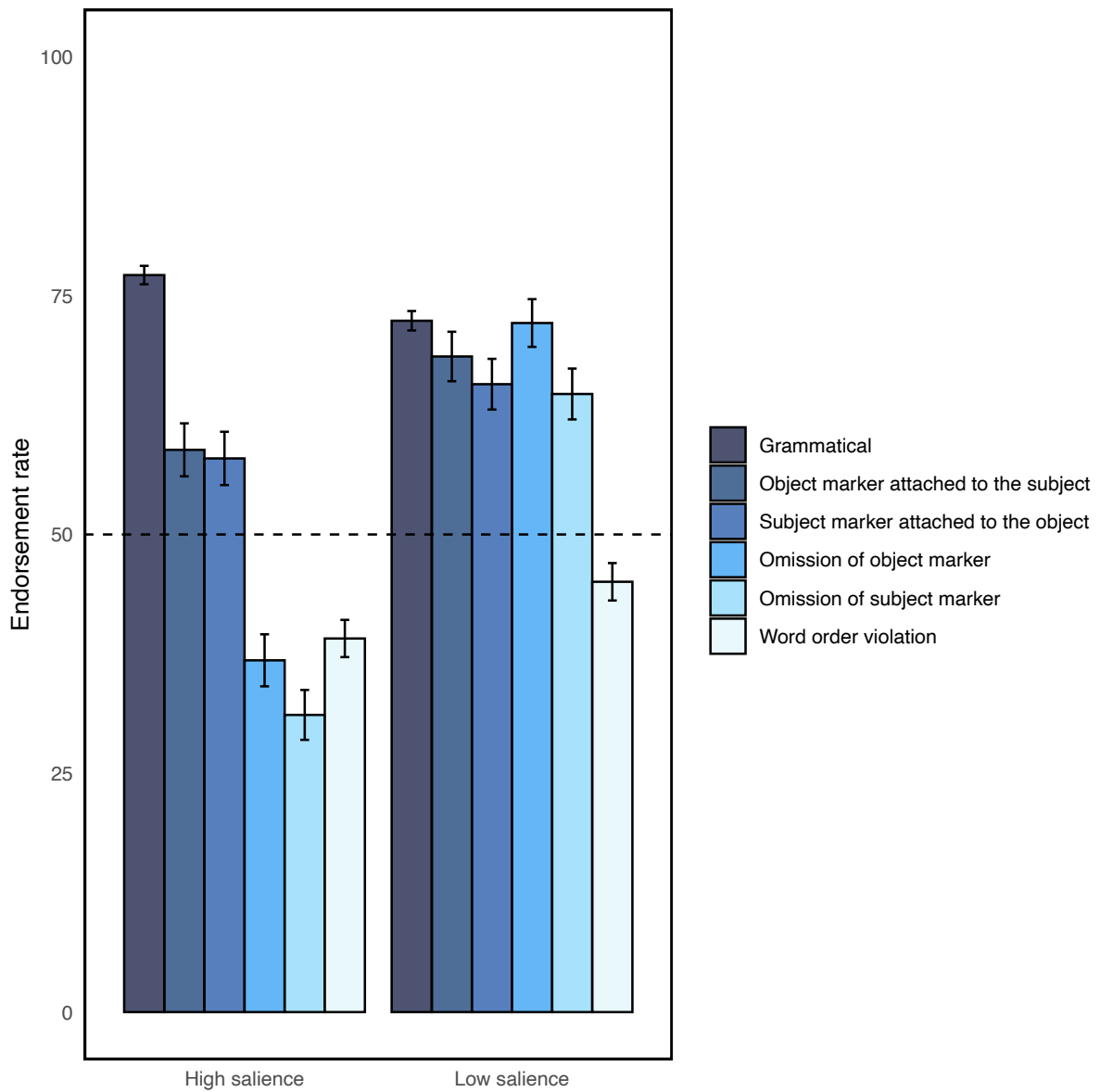
### **Grammaticality judgement task**

Basic descriptive analyses on performance on the grammaticality judgement task (GJT) across salience groups conform to our expectations. Participants in the high-salience group were more accurate overall ( $M = 66.66\%, SE = 0.77$ ) than those in the low-salience group ( $M = 56.07\%, SE = 0.80$ ). The high-salience group was also more likely to endorse grammatical sentences ( $M = 77.15\%, SE = 0.97$ ) and less likely to endorse ungrammatical ones ( $M = 43.83\%, SE = 1.14$ ) compared to the low-salience group (72.37% and 60.20%, respectively).

These patterns are illustrated in Figure 3, which plots endorsement rates by salience group and sentence type. The high-salience group clearly distinguishes between grammatical and ungrammatical structures, with endorsement rates well above chance for grammatical sentences and well below chance for most ungrammatical ones. In contrast, the low-salience group displays a more muted profile, with endorsement rates for ungrammatical sentences often close to or above chance, especially when case markers were omitted. This provides converging evidence that enhanced perceptual salience boosted learners' ability to generalise case-marking patterns.

**Figure 3**

*Endorsement rates for grammatical and ungrammatical sentences in the grammatical judgement task. Error bars show standard error of the mean. Dotted line at 50 per cent shows chance performance.*



To further assess sensitivity while controlling for response bias, we computed d-prime ( $d'$ ) scores for each participant, separately for case marker and word order trials. We then ran

two linear regression analyses on these  $d'$  scores, LM1 And LM2, to test the hypotheses that case marker salience would benefit word order acquisition (H5) and marker syntax (H6) respectively. For LM1, the effect of salience condition on  $d'$  scores in word order trials was not statistically significant,  $\beta = -0.12$ ,  $SE = 0.08$ ,  $t = -1.37$ ,  $p = .173$ , indicating that salience condition did not predict detection of word order violations. A one-sample t-test against chance ( $d' = 0$ ) on the EMMs of LM1 indicated that sensitivity to word order violations was significantly above chance in both the high-salience ( $d' = 0.28$ ,  $p < .001$ ) and low-salience ( $d' = 0.16$ ,  $p < .05$ ) conditions. For LM2, the effect of salience condition on  $d'$  scores in case marker trials was statistically significant,  $\beta = 0.37$ ,  $SE = 0.06$ ,  $t = 6.37$ ,  $p < .001$ , with higher scores in the high-salience group. The model explained approximately 20% of the variance ( $R^2 = .20$ , adjusted  $R^2 = .20$ ). A one-sample t-test against chance ( $d' = 0$ ) on the EMMs of LM2 indicated that sensitivity to case marker violations was significantly above chance in the high-salience condition ( $d' = 0.15$ ,  $p < .001$ ), whereas sensitivity to case marker violations in the low-salience condition was significantly below chance ( $d' = -0.23$ ,  $p < .001$ ).

Finally, we examined correlations between  $d'$  scores and participants' awareness of case marking (from retrospective reports) and LSBQ language background scores. We report Bonferroni-adjusted p-values: For word order  $d'$ , there was no significant correlation with awareness ( $r = -.09$ ,  $p = .26$ ), but a trend of a negative correlation with LSBQ, although not reaching significance ( $r = -.18$ ,  $p = .07$ ). For case marker  $d'$ , there was a weak positive correlation with awareness that did not reach significance after Bonferroni corrections ( $r = .20$ ,  $p = .05$ ), but no significant correlation with LSBQ ( $r = .07$ ,  $p = 1$ ).

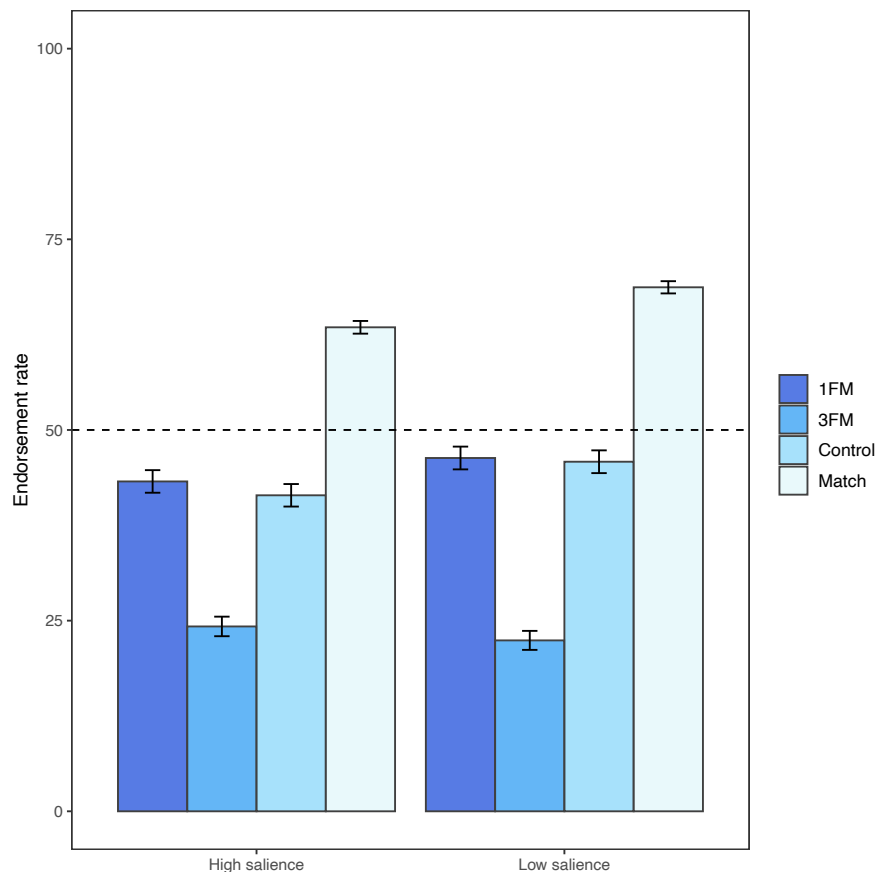
### **Word-Picture Matching Task**

Figure 4 presents endorsement rates of the pseudowords presented in the WPM task by salience group and type of auditory stimuli: the correct form (Match), a 1-feature mismatch (1FM), a 3-feature mismatch (3FM), and a Control item from a different referent.

Accurate phonolexical encoding would be reflected in high endorsement of match forms and low endorsement of all mismatches, particularly those differing by just one phonological feature. Accuracy was comparable across salience conditions, with the high-salience group performing slightly worse ( $M = 63.58\%$ ,  $SE = 0.59$ ) than the low-salience group ( $M = 65.26\%$ ,  $SE = 0.58$ ).

#### Figure 4

*Endorsement rates of the different conditions on the WPM task. Error bars show standard error of the mean. The dotted line at 50% indicates chance performance.*



*Note. For example, for the correct mapping of the pseudoword /ki.ɫu/ (Match), the 1-feature mismatch (1FM) was /kilu/, the 3-feature mismatch (3FM) was /kifu/, and the control (Control) was /diju/.*

GLMER6 was run to explore the phonological detail of the lexical representations formed through cross-situational learning (CSL). The model revealed a significant main effect of mismatch condition in all the specified comparisons. Accuracy for match items was higher than accuracy for controls ( $b = 0.35$ ;  $SE = 0.05$ ;  $z = 7.37$ ;  $p < .001$ ), higher than accuracy for 1FM items ( $b = 0.36$ ;  $SE = 0.06$ ;  $z = 6.04$ ;  $p < .001$ ), but lower than accuracy for 3FM items ( $b = -0.62$ ;  $SE = 0.06$ ;  $z = -11.04$ ;  $p < .001$ ). . However, there was no significant effect of salience ( $b = -0.10$ ;  $SE = 0.07$ ,  $z = -1.45$ ,  $p = .15$ ). To further explore discrimination sensitivity between Match and 1FM items, we computed  $d'$  scores per participant, which showed very limited discriminability and large individual differences ( $M = 0.39$ ,  $min = -0.45$ ,  $max = 1.70$ ).

To more accurately explore the differences in accuracy to pseudowords with different target phonemes and referent types, we computed EMMs for each variable. Regarding the effect of phoneme, EMMs varied across phoneme categories (/e/, /o/, /ɲ/ and /ʎ/), averaged over match, word function, and condition. Accuracy was lowest for /e/ (0.64, 95% CI [0.61, 0.67]), followed by /o/, (0.66, 95% CI [0.62, 0.69]), /ɲ/ (0.66, 95% CI [0.63, 0.70]), and highest for /ʎ/ (0.69, 95% CI [0.66, 0.71]). Pairwise comparisons showed that /e/ yielded significantly lower accuracy than /ʎ/ ( $b = -0.2234$ ,  $SE = 0.0709$ ,  $p < .01$ ). No other pairwise contrasts reached significance. Regarding the effect of referent type, i.e., each word's function during CSL, EMMs showed clear differences among referent type categories (noun, verb, adjective), averaged across match, phoneme, and condition. Accuracy was lowest for nouns (0.62, 95% CI [0.60, 0.64]), followed by adjectives (0.66, 95% CI [0.63, 0.69]), and highest for verbs (0.71, 95% CI [0.68, 0.74]). Pairwise comparisons confirmed that nouns were recognized significantly less accurately than both verbs ( $b = -0.393$ ,  $SE = 0.0635$ ,  $p <$

.001), and adjectives ( $b = -0.171$ ,  $SE = 0.0657$ ,  $p < .05$ ). Verbs were also recognized more accurately than adjectives ( $b = 0.222$ ,  $SE = 0.0739$ ,  $p < .01$ ).

The model explained approximately 1% of the variance in accuracy (marginal pseudo- $R^2 = .05$ ; conditional pseudo- $R^2 = .11$ ), consistent with subtle but reliable effects of phonological distance on performance.

### **AX Discrimination Task**

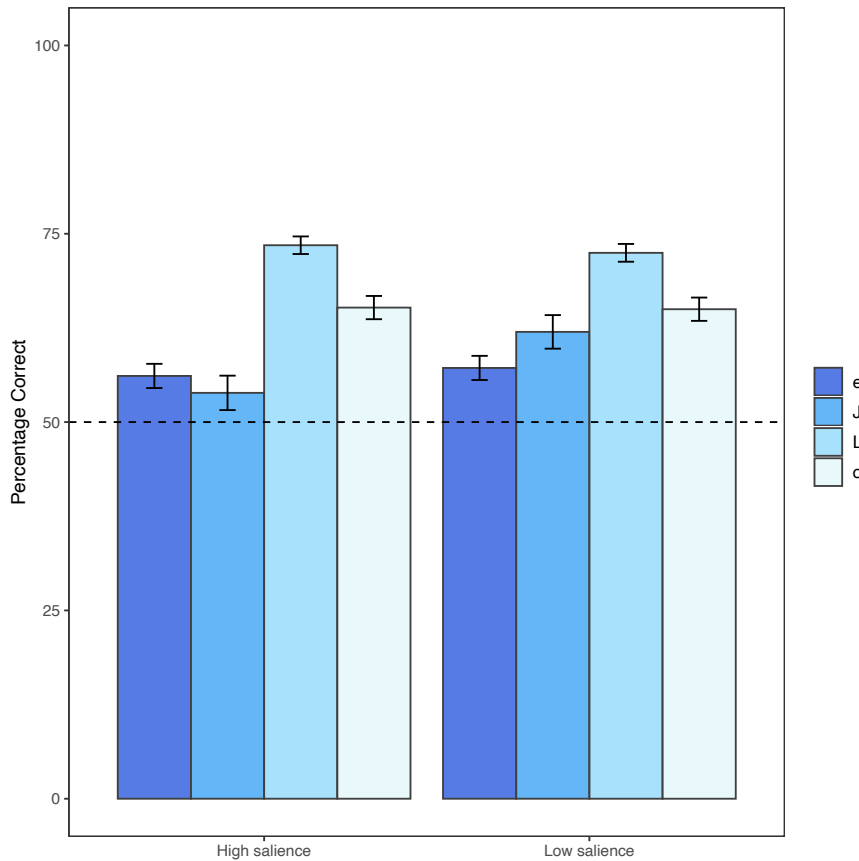
As shown in Figure 5, overall accuracy on the AX discrimination task was comparable across salience groups: high-salience ( $M = 64.61\%$ ,  $SE = 0.77$ ) and low-salience ( $M = 65.47\%$ ,  $SE = 0.77$ ). Across both groups, the /l/-/ʌ/ contrast was most accurately discriminated, followed by /o/-/ɔ/. The /e/-/ɛ/ and /n/-/ŋ/ contrasts yielded lower accuracy overall, though the low-salience group performed somewhat better on the latter contrast than the high-salience group<sup>3</sup>.

### **Figure 5**

*Accuracy in the AX discrimination task for each target phoneme contrast, by salience group. Error bars show standard error of the mean. Dotted line marks chance level at 50%.*

---

<sup>3</sup> An additional glmer model on accuracy on the AX task with salience condition, target phoneme contrast and their interaction as fixed effects including both item-pair and speaker-pair random effects showed no significant between-groups differences in discrimination of the different contrasts.



Note. “e” = /e/-/ɛ/, “J” = /n/-/ɲ/, “L” = /l/-/ʎ/, “o” = /o/-/ɔ/.

We run a Pearson correlation to explore the potential relationship between phoneme discrimination ability and the ability to reject 1-feature mismatches in the WPM task, a key indicator of phonolexical specificity. This revealed a weak but statistically significant relationship between AX accuracy and accuracy in rejecting 1-feature mismatches in the WPM task ( $r = .26, p < .001$ ).

See supplementary materials for a visualisation of this correlation (Figure S6) and a supplementary mixed-effects model (GLMER7) further supporting the conclusion that variability in phonetic perception contributed to observed differences in lexical encoding.

## Discussion

This study examined how adults acquire vocabulary and grammar in an unfamiliar language through cross-situational learning (CSL), focusing in particular on the role of perceptual salience and phonological discrimination. English-speaking participants were exposed to an artificial language modelled on Portuguese, through animated scenes paired with spoken sentences. We followed two main research aims. First, we focused on our pre-registered hypotheses regarding a positive effect of perceptual salience of case markers on case marker comprehension (H1), overall learning (H2), noun (H3) and verb comprehension (H4) and the detection of grammatical violations in word order (H5) and case marker (H6). Then, outside the pre-registration but in line with our pre-registered analyses, we explored whether participants could acquire lexical and morphosyntactic knowledge simultaneously from linguistic input with non-native phonology by tracking cross-situational statistics. Additionally, we explored the phonetic detail of the acquired representations and their relation to phoneme discrimination ability. To test these questions, we assessed vocabulary learning, grammaticality judgment, word–picture matching, and phoneme discrimination. The findings are discussed below.

### **Learning words and grammar in L2 CSL**

This section addresses the general learning patterns, exploring whether learners could acquire vocabulary and grammar in a language with unfamiliar phonology. We hereby evaluate our tentative expectation that participants would successfully learn lexical items and morphosyntactic markers through CSL despite the L2 sound system. Performance on training trials becomes significantly above chance from block 2 onwards in the high-salience group, and from block 3 onwards in the low-salience group. Paired with a significant effect of block in accuracy, this indicates that adults were able to acquire at least part of the novel language through cross-situational statistical learning despite the novel phonology.

Performance on the vocabulary tests was above chance for nouns and verbs in both salience groups by the end of the experiment, indicating partial learning of the corresponding lexical mappings. However, participants failed to acquire adjectives and case markers, as indicated by chance-level performance in these conditions. These results closely mirror those of Rebuschat et al. (2021, Experiment 2), who found a similar pattern—successful learning of nouns and verbs but not adjectives and case markers—even when the artificial language used native phonology. This parallel suggests that the lack of learning for adjectives and markers may reflect a general learning difficulty associated with these categories, rather than one caused by L2 phonology per se.

Our findings stand in partial contrast to those of Escudero (2022), who reported significantly better noun learning when the phonology of the artificial language matched participants' L1. In our case, noun learning was small but evident across groups, and broadly comparable in magnitude to that reported by Rebuschat et al. (2021, Experiment 2), who used the same paradigm with native phonology. This similarity suggests that phonological unfamiliarity may not uniformly hinder lexical acquisition.

Grammar test performance offered a more nuanced picture. On average, participants did not reliably reject sentences with incorrect case marking, indicating that they did not acquire the grammatical function of the case markers. This aligns with the weak performance on case marker vocabulary trials. Nevertheless, participants in the high-salience condition showed above-chance performance in detecting grammatical violations involving marker omission and word order, pointing to some emergent awareness of morphosyntactic regularities, particularly when case markers were perceptually prominent. In contrast, the low-salience group is only above chance in detecting word order violations, albeit marginally. The limited overall grammar learning may, at least partly, be attributable to L1 interference. Given that all learners were L1 English speakers, a VO language without overt case marking,

acquiring an unfamiliar (OV) word order and case morphology may have been a particularly challenging task. Although, in the present study, L1 effects are not directly tested, previous research has documented this kind of L1 interference (Al-Khresheh, 2010; Murakami & Alexopoulou, 2016). Importantly, as we discuss in the following section, our results seem to suggest that increasing perceptual salience facilitates noticing of these novel aspects during early L2 learning, which may have positive consequences for subsequent learning. Taken together, these results suggest that participants were able to acquire key aspects of the vocabulary, and limited aspects of the grammar, particularly when perceptual cues were enhanced.

### **Salience in CSL**

This section addresses the specific question of whether perceptual salience facilitates the acquisition of morphosyntactic structure, and evaluates H1, H2, H3, H4, H5 and H6.

Although the salience effect approached significance in case marker comprehension, both salience groups performed near chance level on case marker vocabulary trials in the final test block. It is therefore possible that case marker learning was simply too limited to produce a clear between-group difference. Thus, our data do not support the prediction that participants in the high-salience condition would show superior comprehension of case morphology, as predicted by H1.

Similarly, we observed no significant salience effect in the comprehension of transitive sentences during the training phase. The absence of knock-on effects may reflect the failure to form reliable form–meaning mappings for the case markers, limiting their utility in sentence parsing. In line with this, salience had no effect on the learning of nouns, verbs, or adjectives. As a result, we partly reject H2, which posited salience-driven benefits to overall learning during exposure. We also reject H3 and H4, which predicted salience-driven benefits in noun and verb learning respectively.

The two salience groups diverged more clearly on the grammar test. Although both groups were able to detect word order violations, only the high-salience group reliably rejected grammatical violations involving case marker omissions. Therefore, we reject H5, which predicted that salience would improve sensitivity to word order grammaticality, but found support for H6, which predicted improved sensitivity to case marker grammaticality. These effects suggest that enhanced salience drew learners' attention to the form and position of the case markers, promoting sensitivity to morphological structure even in the absence of complete functional understanding. This is consistent with the higher reported awareness of case marking suffixes in the high-salience group compared to the low-salience group. Additionally supporting this argument is the trend towards a positive correlation between reported awareness of case marking suffixes and performance in case marker trials. Overall, these findings suggest that increased perceptual salience aids noticing and segmentation more than it establishes functional mappings, at least during early stages of L2 learning.

Consistent with the Noticing Hypothesis (Schmidt, 1990), we interpret the high-salience group's heightened sensitivity to marker omission as evidence of increased structural awareness, albeit weak, likely driven by greater perceptual prominence. This enhanced awareness may reflect noticing of surface features without full mapping to meaning. Indeed, neither group was able to reject incorrect role assignments, indicating that the case markers' grammatical function was not acquired. As mentioned above, one likely explanation is the absence of a direct L1 equivalent: English lacks overt case markers, making it difficult to associate novel morphemes with unfamiliar grammatical functions. This is supported by findings from L2 morphosyntax (Murakami & Alexopoulou, 2016) and artificial language studies (Kenanidis et al., 2023), both of which show greater difficulty with morphosyntactic features lacking L1 analogues. Accordingly, our results suggest only a weak effect of perceptual salience, which may reflect the fact that this notion captures only physical

stimulus properties, whereas broader definitions incorporating psycholinguistic and experiential factors such as redundancy or L1 experience may reveal stronger salience effects (see Knell et al., 2025).

In sum, although perceptual salience did not facilitate vocabulary learning, it played a key role in enhancing learners' sensitivity to morphosyntactic form and structure.

### **Phono-lexical representations**

This section addresses our exploration of the level of phonological detail of the lexical representations formed during CSL, and whether individual differences in phonetic discrimination ability could predict the precision of these representations. We evaluated these questions through the Word–Picture Matching (WPM) task and the AX phoneme discrimination task.

Participants reliably endorsed correct forms (Match) and rejected clearly deviant forms that differed by three phonological features (3FM), indicating that learners acquired word–referent associations with sufficient specificity to detect gross phonological mismatches. However, rejection accuracy dropped substantially for the challenging L2 contrasts (1-feature mismatches), which suggests that the phonological representations formed under CSL are only partially specified. This pattern aligns with previous evidence for phonolexical fuzziness in L2 acquisition (Darcy et al., 2025) and reflects difficulty with encoding subtle contrasts in unfamiliar phonological systems.

Interestingly, rejection of unrelated control pseudowords was also relatively poor, suggesting that form–meaning associations remained weak overall. Such relatively low rejection of control pseudowords, particularly compared to the higher rejection of 3FM pseudowords, could reflect a familiarity effect. Unlike 1FM and 3FM pseudowords, which were created specifically for the WPM task and thus heard for the first time during testing, control pseudowords were previously heard during the CSL task, although linked to different

referents. Given the overall modest vocabulary learning (< 60%), participants may have relied on a coarse familiarity signal rather than precise form-meaning mappings, leading them to endorse familiar-sounding forms even when paired with incorrect referents. This familiarity effect could be attributed to the limited duration of exposure (~40 minutes), or to challenges posed by the unfamiliar phonology, as documented in prior work (Escudero, 2022; Kaushanskaya et al., 2013), or both. Without a native-language control group, we cannot disentangle these two factors.

Regarding the potential relationship between phonetic discrimination ability (AX task) and phonolexical precision (WPM 1FM condition), we observed a significant but weak positive correlation. This finding replicates prior findings that phoneme differentiation accounts for only a small proportion of variance in lexical encoding (Elvin, 2016; Llompart et al., 2025). This suggests that phonolexical fuzziness is only partially rooted in perceptual acuity, and that other individual difference factors, such as working memory, attention, or auditory processing, likely contribute to lexical learning under uncertain conditions.

In summary, our findings indicate that lexical representations acquired through CSL are underspecified, especially for minimal phonological contrasts. We also find a weak but significant link between phonetic discrimination ability and phonolexical specificity.

## **Conclusion**

This study demonstrated that adults are capable of learning both vocabulary and grammatical structure in a novel language with unfamiliar phonology through cross-situational statistical learning. Participants acquired nouns and verbs successfully, and, particularly when exposed to more perceptually salient case markers, showed emerging sensitivity to word order despite the challenges posed by the L2 phonological forms. These results align with previous findings from studies using native phonology (e.g., Rebuschat et

al., 2021), suggesting that CSL mechanisms can still support learning even under more demanding learning conditions.

The role of perceptual salience proved especially relevant for developing sensitivity to grammatical morphemes. Although overall comprehension of case markers remained limited, learners exposed to high-salience markers were more attuned to their omission and placement. This supports the idea that enhancing the perceptual distinctiveness of grammatical morphemes can draw learners' attention to their structural properties, consistent with the predictions of the Noticing Hypothesis (Schmidt, 1990). However, this awareness did not result in robust form–function mappings for the case markers themselves, likely due to the absence of comparable grammatical functions in the learners' native language. This highlights the importance of conceptual alignment between L1 and L2 in the acquisition of morphosyntactic features.

Interestingly, while increased salience facilitated awareness of morphosyntactic structure, it did not enhance lexical learning. There were no measurable benefits for the learning of nouns, verbs, or adjectives. Nonetheless, we take this to suggest that perceptual factors can influence syntactic awareness even in the absence of full morphological comprehension.

Turning to the nature of the lexical representations acquired, we found evidence that participants encoded word forms imprecisely: while they readily accepted correct forms and rejected clearly mismatching lures, they struggled with minimally different distractors. This pattern suggests that phonological representations were underspecified, a finding consistent with previous accounts of lexical fuzziness in L2 word learning. Although phonetic discrimination ability contributed to this variability, the relationship was relatively weak, pointing to the likely involvement of other relevant factors.

In sum, our findings underscore the effectiveness of CSL for supporting early L2 learning and highlight the important, but limited, role of perceptual salience in facilitating morphosyntactic awareness. They also reveal the fragile and fuzzy nature of early lexical representations when learning occurs under conditions of phonological unfamiliarity. Future research should build on these findings by examining how phonological, conceptual, and perceptual factors interact to shape the trajectory of L2 development across learning contexts.

Data availability: The data that support the findings of this study are openly available in OSF at [https://osf.io/zm9q6/?view\\_only=2aba48140aa54f52a008410a23af2c2f](https://osf.io/zm9q6/?view_only=2aba48140aa54f52a008410a23af2c2f) (provisional (anonymised) link for peer review).

## References

- Anderson, J. A., Mak, L., Keyvani Chahi, A., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior research methods*, *50*, 250-263. <https://doi.org/10.3758/s13428-017-0867-9>
- Audacity Team (2024). Audacity®: Free Audio Editor and Recorder [Computer program]. Version 3.2.3 retrieved August 5<sup>th</sup> 2024 from <https://www.audacityteam.org/download/>
- Al-Khresheh, M. (2010). Interlingual interference in the English language word order structure of Jordanian EFL learners. *European journal of social sciences*, *16*(1), 106-113.
- Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis Model of the L2 Lexical Representation. *Bilingualism: Language and Cognition*, *25*(2), 185–201. <https://doi.org/10.1017/S1366728921000250>
- Boswijk, V., & Coler, M. (2020). What is salience?. *Open Linguistics*, *6*(1), 713-722. <https://doi.org/10.1515/opli-2020-0042>
- Clements, G. N. (1990). The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, *1*, 283-333.
- Correia, S., Rato, A., Ge, Y., Fernandes, J. D., Kachlicka, M., Saito, K., & Rebuschat, P. (2025). Effects of phonetic training and cognitive aptitude on the perception and production of non-native speech contrasts. *Studies in Second Language Acquisition*, *47*(1), 440-457. doi:10.1017/S0272263124000548
- Darcy, I., Llompart, M., Hayes-Harb, R., Mora, J. C., Adrian, M., Cook, S., & Ernestus, M. (2025). Phonological processing and the L2 mental lexicon: Looking back and moving forward. *Studies in Second Language Acquisition*, *47*(1), 361–387. doi:10.1017/S0272263124000482

- Ellis, N. C. (2017). Salience in usage-based SLA. In S. Gass, P. Spinner, & J. Behney. (Eds.), *Salience in second language acquisition* (pp. 39–58). New York, NY: Routledge.
- Elvin, J. D. (2016). *The role of the native language in non-native perception and spoken word recognition: English vs. Spanish learners of Portuguese* [PhD Thesis, Western Sydney University (Australia)].
- Escudero, P., Smit, E. A., & Mulak, K. E. (2022). Explaining L2 Lexical Learning in Multiple Scenarios: Cross-Situational Word Learning in L1 Mandarin L2 English Speakers. *Brain Sciences*, 12(12), 1618. <https://doi.org/10.3390/brainsci12121618>
- Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2017). Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking. *Cognitive Science*, 41(2), 416–446. <https://doi.org/10.1111/cogs.12346>
- Ge, Y., Correia, S., Fernandes, J. D., Rato, A., & Rebuschat, P. (2024). Does Phonetic Training Benefit Word Learning?. <https://doi.org/10.31234/osf.io/5zspu> [Preprint].
- Ge, Y., Correia, S., Lee, Y.-W., Jin, Z., Rothman, J., & Rebuschat, P. (2025a). Statistical learning of foreign language words in younger and older adults. *Bilingualism: Language and Cognition*, 28(3), 716–727. doi:10.1017/S1366728924000907
- Ge, Y., Monaghan, P., & Rebuschat, P. (2025b). The role of phonology in non-native word learning: Evidence from cross-situational statistical learning. *Bilingualism: Language and Cognition*, 28(1), 15–30. doi:10.1017/S1366728923000986
- Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the “natural order of L2 morpheme acquisition” in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1), 1-50. <https://doi.org/10.1111/1467-9922.00147>
- Haznedar, B. (2008). Persistent problems with case morphology in L2 acquisition. In C. Lle (Ed.) *Interfaces in multilingualism: Acquisition and representation* (pp. 179-206). John Benjamins Publishing Company.

Kaushanskaya, M., Yoo, J., & Van Hecke, S. (2013). Word learning in adults with second-language experience: Effects of phonological and referent familiarity. *Journal of Speech, Language, and Hearing Research*, 56(2), 667-678. [https://doi.org/10.1044/1092-4388\(2012/11-0084\)](https://doi.org/10.1044/1092-4388(2012/11-0084))

Kenanidis, P., Dąbrowska, E., Llompart, M., & Pili-Moss, D. (2023). Can adults learn L2 grammar after prolonged exposure under incidental conditions? *PLOS ONE*, 18(7), e0288989. <https://doi.org/10.1371/journal.pone.0288989>

Kenanidis, P., Llompart, M., Fernández Santos, S., & Dąbrowska, E. (2024). Redundancy can hinder adult L2 grammar learning: evidence from case markers of varying salience levels. *Frontiers in Psychology*, 15, 1368080. <https://doi.org/10.3389/fpsyg.2024.1368080>

Knell, G., Cipitria, S., De Cuypere, L., Housen, A., & Struys, E. (2025). Stand-out: A systematic review of the role of salience in second language acquisition. *Journal of the European Second Language Association*, 9(1). <https://doi.org/10.22599/jesla.131>

Llompart, M. (2021). Phonetic categorization ability and vocabulary size contribute to the encoding of difficult second-language phonological contrasts into the lexicon. *Bilingualism: Language and cognition*, 24(3), 481-496. <https://doi.org/10.1017/S1366728920000656>

Llompart, M., Gorba, C., & Prieto, P. (2025). Revisiting the link between second-language sound identification and word recognition with an eye on methodological similarity. *Studies in Second Language Acquisition*, 1-13. <https://doi.org/10.1017/S0272263125101113>

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as Gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive science*, 39(5), 1099-1112. <https://doi.org/10.1111/cogs.12186>

Monaghan, P., Ruiz, S., & Rebuschat, P. (2021). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local

and global effects on acquisition. *Second Language Research*, 37(2), 261-289.

<https://doi.org/10.1177/0267658320927741>

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3), 365-401. <https://doi.org/10.1017/S0272263115000352>

Plonsky, L. (2023). Sampling and generalisability in Lx research: A second-order synthesis. *Languages*, 8(1), 75. <https://doi.org/10.3390/languages8010075>

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT press.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, 63(3), 595-626. <https://doi.org/10.1111/lang.12010>

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition*, 206, 104475.

<https://doi.org/10.1016/j.cognition.2020.104475>

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied linguistics*, 11(2), 129-158. <https://doi.org/10.1093/applin/11.2.129>

Scott, R. M., & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, 122(2), 163-180.

<https://doi.org/10.1016/j.cognition.2011.10.010>

Selkirk, E. (1984). On the major class features and syllable theory. *Language sound structure*.

Simoens, H., Housen, A., & De Cuypere, L. (2017). The effect of perceptual salience on processing L2 inflectional morphology. In S. Gass, P. Spinner, & J. Behney. (Eds.), *Salience in second language acquisition* (pp. 107-130). Routledge.

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, 70(S2), 221-254. <https://doi.org/10.1111/lang.12395>

Williams, J. N. & Rebuschat, P. (2023). Implicit learning and SLA: A cognitive psychology perspective. In A. Godfroid and H. Hopp (Eds), *The Routledge Handbook of Second Language Acquisition and Psycholinguistics* (pp. 281-293). Routledge.

Yu, C., & Smith, L. B. (2008). Statistical cross-situational learning in adults and infants. In *Proceedings of the 32th Annual Boston University Conference on Language Development* (pp. 562-573). Somerville, MA: Cascadilla Press.

Zhu, L., Rebuschat, P., Nixon, J. S., & Monaghan, P. (2025). Learning morphology from cross-situational statistics. *Studies in Second Language Acquisition*, 1-27. <https://doi.org/10.1017/S027226312510106X>

## Supplementary Materials

### Pseudowords in each task

**Table S1**

*The pseudo-words used in the CSL task.*

Categories	Target segments	Pseudowords
Consonants	/k/	/kiɮu/
		/suɮu/
		/tiɮu/
	/p/	/muɮu/
		/diɾu/
		/piɾu/
Vowels	/e/	/fuɾu/
		/detu/
		/kepu/
	/o/	/pemu/
		/megu/
		/sodu/
		/togu/
		/fobu/

**Table S2***The pseudo-words used in the WPM task.*

Target phoneme	Match	1 feature mismatch	3 features mismatch	Control (unrelated)
/k/	/kiɮu/	/kɪlu/	/kɪfu/	/dɪɲu/
	/suɮu/	/sulu/	/sufu/	/tɔgu/
	/tiɮu/	/tilu/	/tɪfu/	/detu/
	/muɮu/	/mulu/	/mufu/	/kepu/
/ɲ/	/dɪɲu/	/dɪnu/	/dɪku/	/pemu/
	/pɪɲu/	/pɪnu/	/pɪku/	/megu/
	/fɪɲu/	/fɪnu/	/fɪku/	/sodu/
/e/	/detu/	/detu/	/dutu/	/tiɮu/
	/kepu/	/kepu/	/kupu/	/muɮu/
	/pemu/	/pemu/	/pumu/	/fobu/
	/megu/	/megu/	/mugu/	/pɪɲu/
/o/	/sodu/	/sɔdu/	/sidu/	/fɪɲu/
	/tɔgu/	/tɔgu/	/tɪgu/	/kiɮu/
	/fobu/	/fɔbu/	/fɪbu/	/suɮu/

**Table S3**

*The phonetic contrasts used in the WPM task.*

		Phoneme (match)	1FM	3FM
		<i>/ʎ/</i>	<i>/l/</i>	<i>/ʃ/</i>
Phonological features	Place of articulation	palatal	alveolar	post-alveolar
	Manner of articulation	liquid	liquid	fricative
	Voicing	voiced	voiced	voiceless
		<i>/ŋ/</i>	<i>/n/</i>	<i>/k/</i>
Phonological features	Place of articulation	palatal	alveolar	velar
	Manner of articulation	nasal	nasal	stop
	Voicing	voiced	voiced	voiceless
		<i>/e/</i>	<i>/ɛ/</i>	<i>/u/</i>
Phonological features	Height	close-mid	open-mid	close
	Tongue position (horizontal)	front	front	back
	Lip rounding	unrounded	unrounded	rounded
		<i>/o/</i>	<i>/ɔ/</i>	<i>/i/</i>
Phonological features	Height	close-mid	open-mid	close
	Tongue position (horizontal)	back	back	front
	Lip rounding	rounded	rounded	unrounded

*Note.* The differences of conditions 1FM and 3FM with the Match condition in terms of phonological features are shown in red.

**Table S4***The pseudo-word pairs used in the AX discrimination task*

Category	Contrast	Pseudo-words
Consonants	/l/-/ʎ/	/dulu/-/duʎu/
		/silu/-/siʎu/
		/falu/-/faʎu/
	/n/-/ɲ/	/cunu/-/cuɲu/
		/funu/-/fuɲu/
		/pinu/-/piɲu/
Vowels	/e/-/ɛ/	/tedu/-/tɛdu/
		/semu/-/sɛmu/
		/mepu/-/mɛpu/
	/o/-/ɔ/	/dodu/-/dɔdu/
		/sodu/-/sɔdu/
		/tosu/-/tɔsu/

**Table S5**

*Wald z-tests for estimated marginal means on performance for each salience group at each block in all training trials of the CSL task.*

<b>Salience group</b>	<b>Block</b>	<b>Estimated probability</b>	<b>SE</b>	<b>z</b>	<b>p (Holm adjusted)</b>
Low	1	0.49	0.01	-1.09	.55
High		0.51	0.01	0.83	.55
Low	2	0.51	0.01	1.35	.53
High		0.53	0.01	3.01	< .05
Low	3	0.54	0.01	3.07	< .05
High		0.56	0.01	4.33	< .001
Low	4	0.57	0.02	4.01	< .001
High		0.58	0.02	4.91	< .001
Low	5	0.59	0.02	4.50	< .001
High		0.60	0.02	5.15	< .001
Low	6	0.61	0.02	4.77	< .001
High		0.63	0.02	5.25	< .001
Low	7	0.64	0.03	4.94	< .001
High		0.65	0.03	5.29	< .001
Low	8	0.66	0.03	5.05	< .001
High		0.67	0.03	5.30	< .001

---

Low	9	0.67	0.03	5.12	< .001
High		0.69	0.03	5.30	< .001

---

Low	10	0.70	0.03	5.17	< .001
High		0.71	0.03	5.30	< .001

---

Low	11	0.73	0.04	5.21	< .001
High		0.73	0.04	5.30	< .001

---

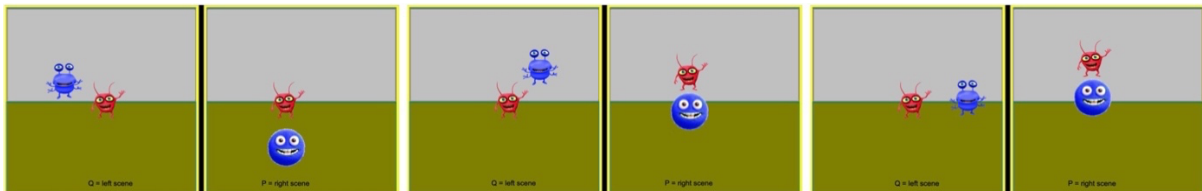
Low	12	0.75	0.04	5.25	< .001
High		0.75	0.04	5.28	< .001

---

## Visualization of trial types

### Figure S1

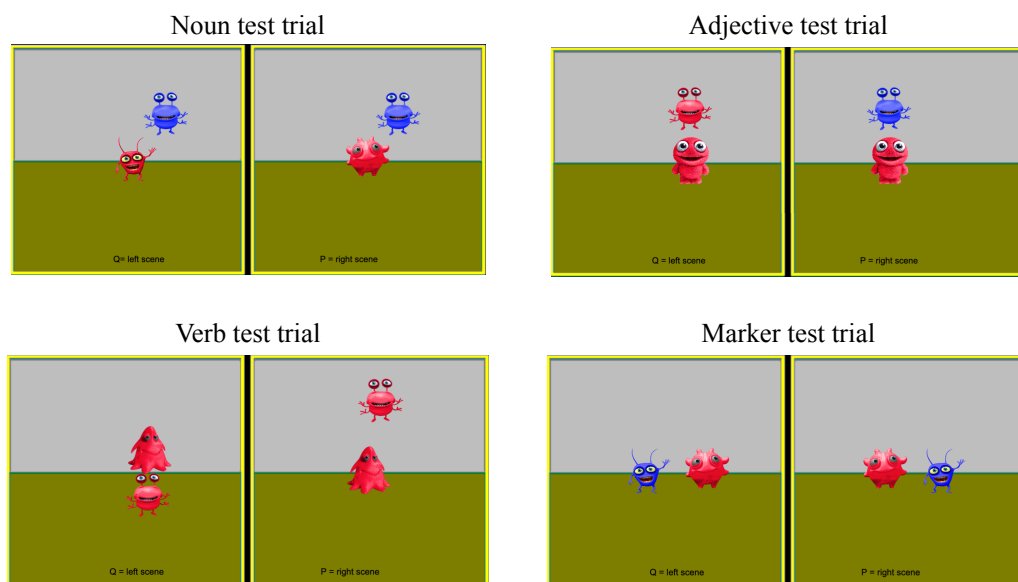
*Static images of the animations shown during one single exposure trial at 3 consecutive points in time*



*Note.* The left-side animation shows the action of ‘jumping’ and the right-side animation shows the action of ‘lifting’. The animations were shown on a loop, so that they re-played until an answer was given.

### Figure S2

*Static images of scene pairs from the four types of vocabulary test trials*

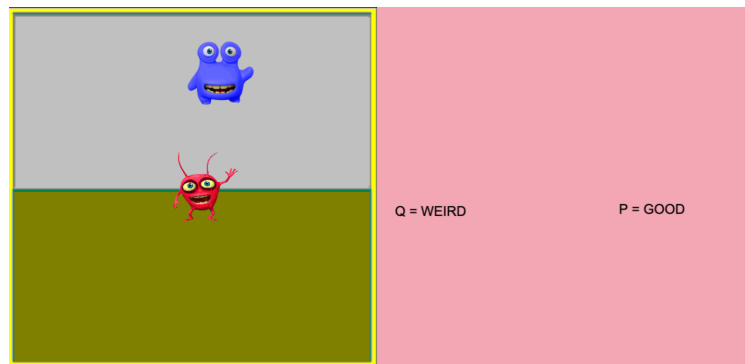


*Note.* In the **noun test trial**, the same action (*jumping over*) is performed by identical agent aliens, but with different patients (same color but different alien). The **adjective test trial** shows the same action (*lifting*) by identical agent aliens, but with patients of different colors (same alien but different color). In the **verb**

**test trial**, identical aliens with the same roles perform different actions (*lifting* on the left, *jumping over* on the right). The **marker test trial** shows the same action by the same aliens, but with agent and patient roles reversed between scenes.

### Figure S3

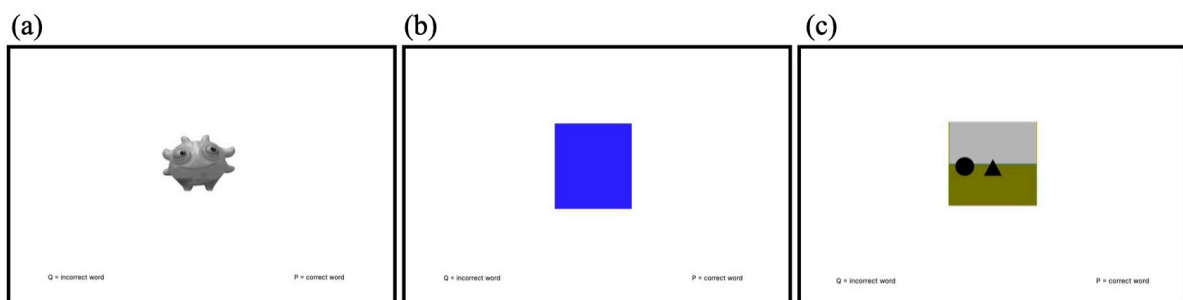
*Static image of a grammaticality judgement task trial*



*Note.* The scene always appeared on the left side and matched the (grammatical or ungrammatical) spoken sentence.

### Figure S4

*The three possible types of visual referents in the WPM task*

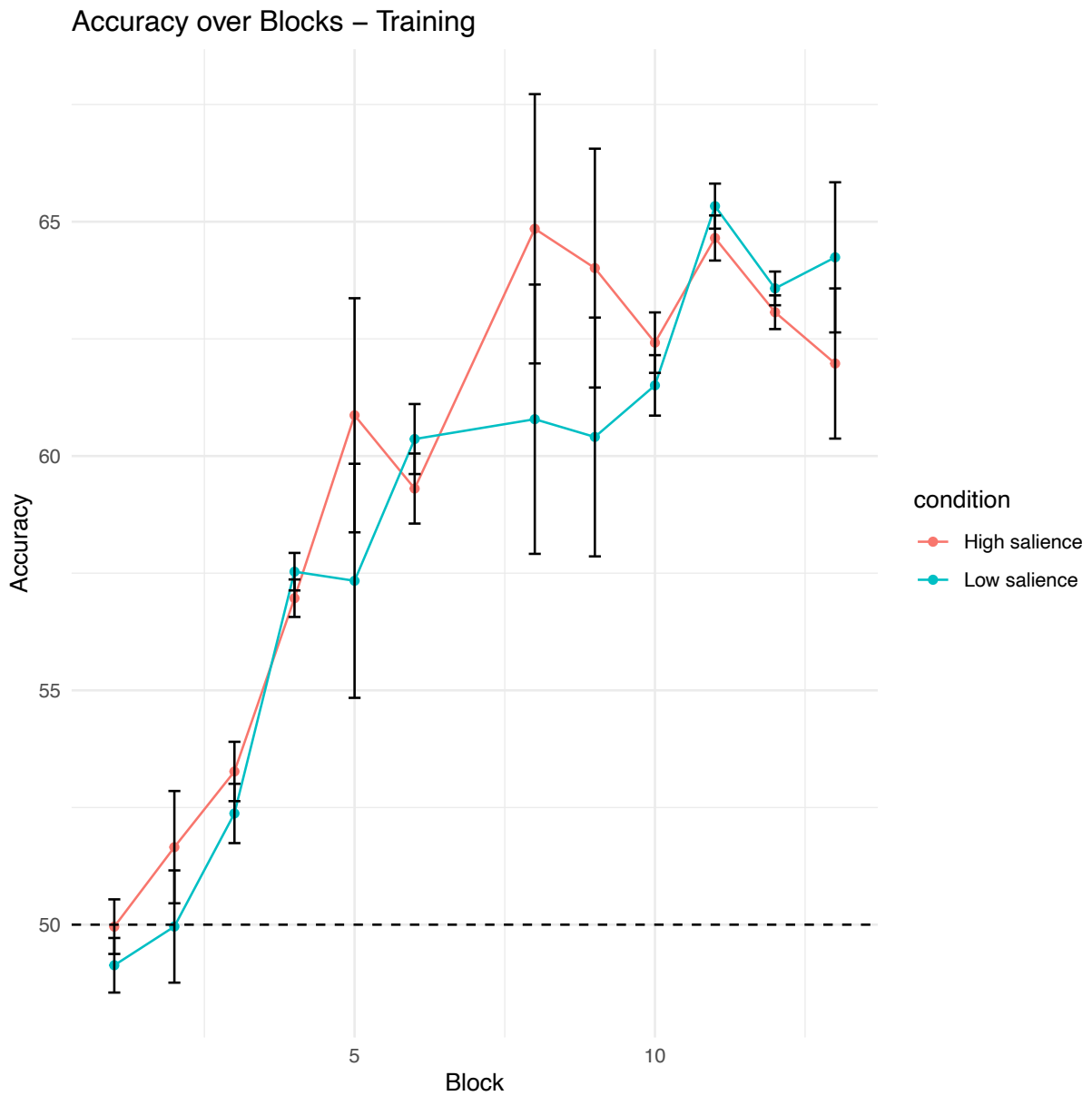


*Note.* Participants were presented with either a black and white alien character (a), a color (b) or an animation of an action performed by black shapes (c) and heard a pseudoword that could match the learned association for the alien, color or action respectively or not.

## Additional visualisation of results

**Figure S5**

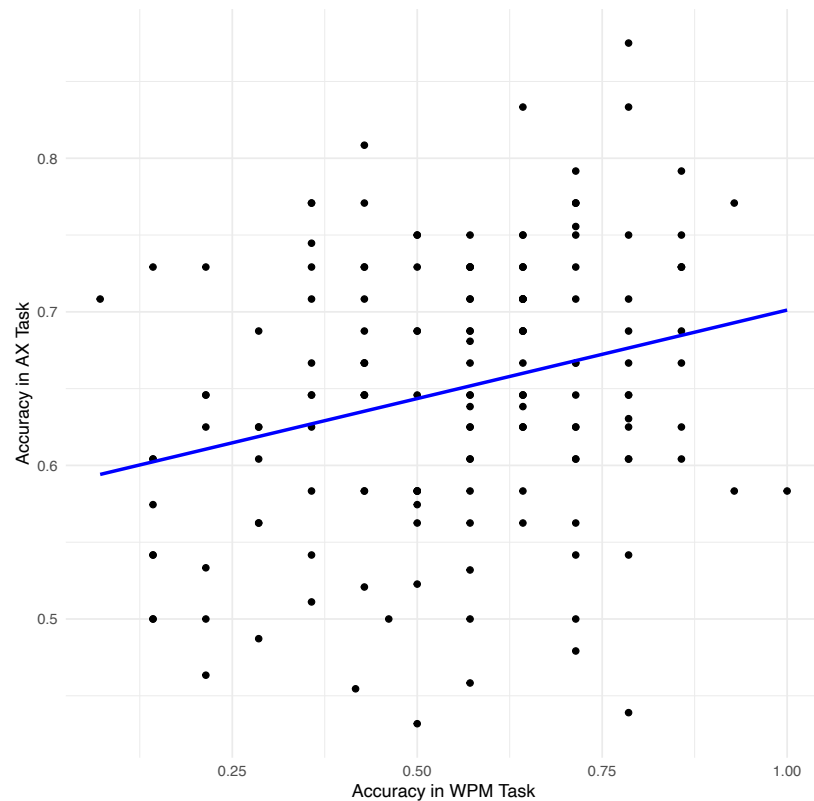
*Performance accuracy on exposure trials*



*Note.* The error bar shows standard error of the mean. The graph is scaled from 40% to 70% accuracy for easier visualization. Dotted line at 50 per cent shows chance performance.

## Figure S6

*Scatterplot of individual accuracies in the WPM task by individual accuracies in the AX task*



## Statistical analysis

### *Additional analyses*

To further explore possible learning effects, we conducted one-way ANOVAs with block as a within-subjects variable for each vocabulary test type and salience group separately. For adjectives, case markers and nouns there was no significant learning effect in either group: adjectives high salience:  $F(2.90, 214.75) = 0.65, p = .580, \eta_p^2 = .006$ , adjectives low salience:  $F(2.81, 213.19) = 2.04, p = .113, \eta_p^2 = .021$ , case markers high salience:  $F(2.84, 213.00) = 1.42, p = .238, \eta_p^2 = .014$ , case markers low salience:  $F(2.83, 217.73) = 0.72, p = .534, \eta_p^2 = .007$ , nouns high salience:  $F(2.84, 210.39) = 0.35, p = .781, \eta_p^2 = .003$ , nouns low salience:

$F(2.73, 210,38) = 2.33, p = .081, \eta_p^2 = .022$ . For verbs, there was no significant learning effect in the high salience group,  $F(2.85, 213.94) = 1.61, p = .189, \eta_p^2 = .012$ , but there was a significant learning effect in the low salience group,  $F(2.91, 224.12) = 5.27, p < .01, \eta_p^2 = .036$ . Polynomial contrasts confirmed that linear trend was significant, indicating a linear increase in accuracy across blocks,  $t(77) = 3.40, p = .001$ . The quadratic trend was not significant,  $p = 1$ , nor was the cubic trend,  $p = .101$ .

**GLMER7.** To further explore the relationship between acquired phonemic representations and phonetic discrimination ability, we ran another mixed-effects logistic regression using data from the 1-feature mismatch condition in the word-picture matching task and from the AX discrimination task. In this model, overall task accuracy was the categorical dependent variable (1 = correct; 0 = incorrect) and task (word-picture matching task; AX discrimination task), salience condition (high; low), target phoneme (/k/; /n/, /e/; /o/) and the interaction between task and phoneme were the categorical fixed effects. Task was contrast-coded so that the WPM was coded as 0.5 and the AX was coded as -0.5. The random effects structure of the model included random intercepts by item and participant and random slopes by salience over item and by task over participant. Random slopes by phoneme over participant were not included due to near-perfect correlations with the random intercepts, indicating no added value to the model fit. The results showed that none of the fixed effects –task, phoneme, salience, or the interaction between task and phoneme– were statistically significant (task:  $b = -0.28; SE = 0.29, z = -0.95; p = .34$ ; phoneme:  $b = 0.08; SE = 0.26, z = 0.32; p = .75$ ; salience:  $b = -0.04; SE = 0.08, z = -0.54; p = .59$ ; task\*phoneme:  $b = -0.69; SE = 0.52, z = -1.32; p = .19$ ). The model explained 1% of the variance observed (marginal pseudo- $R^2 = .02$ ; conditional pseudo- $R^2 = .38$ ).

### ***Model specification***

All GLMMs were implemented using the `glmer` function with the optimizer set to “bobyqa” via the `glmer` control options. As with all logistic regressions reported in this study, Johnson’s (2014) conditional and marginal pseudo- $R^2$  values were calculated using the `r.squaredGLMM` function from the `MuMIn` package (version 1.47.5; Barton & Barton, 2015). Assumption diagnostics were performed using the `DHARMA` package (Hartig & Hartig, 2017).

### ***Model selection details***

***GLMER1 (H1)***. Dependent variable was trial-level accuracy on case marker test trials in the final block of the CSL task. Random slopes for condition by item were excluded due to the emergence of a perfect negative correlation between the random slope and intercept, indicating potential model overfitting. Moreover, the inclusion of these slopes did not significantly improve model fit,  $\chi^2(2) = 3.97$ ,  $p = .14$ .

***GLMER2 (H2)***. Dependent variable was trial-level accuracy across all test trials (case marker, noun, verb, and adjective). A maximal random-effects structure including by-item random slopes for condition resulted in overparameterization, as indicated by near-zero variance estimates and unstable intercept–slope correlations. Likelihood-ratio model comparison showed that including random slopes for condition over items did not improve model fit,  $\chi^2(9) = 7.79$ ,  $p = .56$ , and the simpler random-effects structure was therefore retained.

Model diagnostics were assessed using simulation-based residual analyses (`DHARMA`). The nonparametric dispersion test indicated no evidence of over- or underdispersion (dispersion = 1.02,  $p = .37$ ), and the outlier test did not reveal an excess of extreme residuals ( $p = .95$ ). The Kolmogorov–Smirnov test for residual uniformity was significant ( $D = 0.013$ ,  $p < .001$ ); however, given the large number of observations, this

deviation was small in magnitude and is unlikely to reflect a substantive model violation.

Overall, diagnostic checks suggested an adequate model fit, though results should be interpreted with appropriate caution.

**GLMER3–5 (H3, H4).** Separate models for dependent variable: accuracy in last test block for (1) nouns (2) verbs, and (3) adjectives. Random slopes for condition by item were excluded from all three models due to perfect negative correlations with the random intercepts, indicative of overfitting. The addition of these slopes did not significantly improve model fits (nouns:  $\chi^2(2) = 0.49$ ,  $p = .78$ ; verbs:  $\chi^2(2) = 0$ ,  $p = 1$ ; adjectives:  $\chi^2(2) = 2.63$ ,  $p = .27$ ).

**GLMER6 (WPM Task).** Dependent variable: accuracy on all trials of the Word-Picture Matching (WPM)Task. Random slopes for phoneme over participant were not included as their exclusion did not significantly improve model fit and increased model complexity. A likelihood-ratio test comparing the model with and without the random slope showed no evidence for improved fit,  $\chi^2(9) = 8.07$ ,  $p = .53$ .

## Retrospective verbal reports

As a final step, we ask you some questions about your experience in this experiment.

### Question 1:

During the first part of this study, you saw two scenes and heard one sentence. Your task was to choose which scene the sentence referred to. How did you decide which scene the sentence described? Did you just guess throughout the experiment or did you follow any particular strategies? If so, what strategies did you follow?

### Question 2:

Do you think the way you made decisions on the scenes changed throughout the experiment?

### Question 3:

If you knew the names for one of the scenes already, and they did not match the sentence you heard, how did you use that information?

### Question 4:

Did you notice any particular patterns or rules while performing the task?

### Question 5:

Did you notice that some words ended in “-L”? If so, when did you notice (e.g. before or after the break)?

### Question 6:

Did you notice that some words ended in “-R”? If so, when did you notice (e.g. before or after the break)?

### Question 7:

How easy was it to notice these common endings (-L and -R)? Please slide the circle below to indicate the difficulty from “very difficult”, meaning that you barely noticed them to “very easy”, meaning that you noticed them early in the experiment.

*Slider scale from Very difficult to Very easy*

### Question 8:

Do you think words with these common endings (-L and -R) were easier to learn? Please slide the circle below to indicate the difficulty of learning words with endings (as compared to words without endings).

*Slider scale from More difficult to Easier*

**Question 9:**

What real language did you think our alien language sounded like?

**Question 10:**

What do you think was the aim of this study?

## References

Barton, K., & Barton, M. K. (2015). *Package 'mumin'*. R package version 1.47.5.

Hartig, F., & Hartig, M. F. (2017). *Package 'dharma'*. *R package*.