UNIVERSITY of York

This is a repository copy of *The effect of verb surprisal on the acquisition of second language syntactic structures in adults:An artificial language learning study*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/200906/

Version: Published Version

Article:

Bovolenta, Giulia orcid.org/0000-0003-4139-6446 and Marsden, Emma orcid.org/0000-0003-4086-5765 (2023) The effect of verb surprisal on the acquisition of second language syntactic structures in adults:An artificial language learning study. Applied Psycholinguistics. ISSN: 1469-1817

https://doi.org/10.1017/S0142716423000334

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

ORIGINAL ARTICLE

The effect of verb surprisal on the acquisition of second language syntactic structures in adults: An artificial language learning study

Giulia Bovolenta[1] and Emma Marsden[2]

[1]School of Arts, Culture and Language, Bangor University, Bangor, UK and [2]Department of Education, University of York, York, UK
Corresponding author: Giulia Bovolenta; Email: g.bovolenta@bangor.ac.uk

Abstract

Inverse probability adaptation effects (the finding that encountering a verb in an unexpected structure increases long-term priming for that structure) have been observed in both L1 and L2 speakers. However, participants in these studies all had established representations of the syntactic structures to be primed. It therefore remains an open question whether inverse probability adaptation effects could take place with newly encountered L2 structures. In a pre-registered experiment, we exposed participants ($n = 84$) to an artificial language with active and passive constructions. Training on Day 1 established expectations for specific co-occurrence patterns between verbs and structures. On Day 2, established patterns were violated for the surprisal group ($n = 42$), but not for the control group ($n = 42$). We observed no immediate priming effects from exposure to high-surprisal items. On Day 3, however, we observed an effect of input variation on comprehension of verb meaning in an auditory grammaticality judgment task. The surprisal group showed higher accuracy for passive structures in both tasks, suggesting that experiencing variation during learning had promoted the recognition of optionality in the target language.

Introduction

Prediction and prediction error are topics of growing interest in the field of second language (L2) acquisition studies (Bovolenta & Marsden, 2021b). There is evidence to suggest that formulating expectations which are not met, broadly speaking, can enhance learning of new input. For instance, new words can be better learned when they are unexpected (Gambi et al., 2021; Stahl & Feigenson, 2017), due to a phenomenon known as one-shot declarative learning which is found in a variety of domains besides vocabulary learning (De Loof et al., 2018; Greve et al., 2017).

Another potential mechanism by which unmet expectations can enhance learning, this time specific to language, is implicit error-based learning (Chang et al., 2006). This mechanism, which forms the theoretical background for the current study, posits a unified mechanism for language processing and learning that is driven by prediction error. The hypothesis is that learners are constantly formulating expectations about upcoming linguistic input based on their knowledge of the statistical distribution of the language, and when those expectations are not met, they revise their expectations accordingly (in a manner proportional to the magnitude of the prediction error), which amounts to learning.

Computational models implementing implicit error-based learning can reproduce behavioral findings from both first language (L1) acquisition (Hirsh-Pasek & Golinkoff, 1996; Naigles, 1990) and processing, specifically structural priming in adults (Chang et al., 2006). Additional behavioral evidence in favor of implicit error-based learning accounts comes from inverse frequency priming: the finding that syntactic priming effects are stronger when the structure to be primed is encountered in an unexpected context, normally a verb that is not frequently used with that structure (Bernolet & Hartsuiker, 2010).

Inverse frequency priming has been observed in both L1 and L2 speakers (Fazekas et al., 2020; Jackson & Hopp, 2020; Montero-Melis & Jaeger, 2020) which, insofar as the phenomenon can be taken as an indication of implicit error-based learning, suggests that error-based learning is operating in L2 as well as L1. However, even the L2 speakers involved in these studies already had existing L2 representations of the target structure at the time of testing. Therefore, while such findings provide valuable information on L2 processing, there is still limited empirical evidence on whether prediction error can play a role in the L2 *learning* process—specifically, the establishment of new representations, which is the gap addressed by our study.

The aim of the present study was to investigate whether implicit error-based learning can operate at the earliest stages of L2 learning. The behavioral phenomenon we chose to investigate is inverse frequency priming, which, if observed, would suggest that an implicit error-based learning is at play. We created an artificial language study in which we manipulated verb surprisal by varying the statistical patterns of co-occurrence between specific lexical verbs and syntactic constructions. Our research question was whether experiencing higher verb surprisal would induce inverse frequency priming effects, even at the earliest stages of exposure to a new language. Below, we describe the theoretical background and existing evidence on error-based learning with specific reference to L2 acquisition.

Prediction error in language processing and learning

When we are listening to language, we are constantly and automatically forming predictions about what is coming next (Kuperberg & Jaeger, 2016). Computational models of language processing (Chang et al., 2006; Elman, 1990) suggest that prediction mechanisms may not only be helpful for comprehension but may be implicated in language learning, too. In these models, prediction error is suggested as the link between processing and learning: when predictions are disconfirmed, the model adjusts its expectations, gradually adapting to the statistical distribution of

the language. The source of prediction error in these models is operationalized as *surprisal*, which refers to the likelihood of a specific word given the preceding context (Hale, 2001; Levy, 2008). Word-by-word surprisal from these models correlates with language processing in humans, measured by reading times (Frank, 2013; Frank & Hoeks, 2019; Goodkind & Bicknell, 2018; Monsalve et al., 2012; Van Schijndel & Linzen, 2018), N400 amplitudes during EEG (Frank et al., 2013, 2015), and MEG responses (Wehbe et al., 2014), suggesting that humans are sensitive to the same statistical properties of language (surprisal) which generate prediction error in computational models.

A particularly influential model of language processing and acquisition which is based on prediction error is the Dual-Path model (Chang et al., 2006). This connectionist model is based on a recurrent neural network trained on next-word prediction. As the model encounters more sentences, it gradually improves its predictions by adjusting its weights based on the magnitude of the prediction error, i.e., the discrepancy between predicted and actual input (Chang et al., 2006). This model can reproduce data from child language acquisition (Hirsh-Pasek & Golinkoff, 1996; Naigles, 1990) and from structural priming in adults (Chang et al., 2006). The Dual-Path model's ability to reproduce phenomena from L1 acquisition and processing suggests that these may be driven by prediction error: as we encounter unexpected (high-surprisal) input, we update our representations to match that input, which amounts to learning. Therefore, there is growing interest in the role that prediction error may play in first language acquisition (Fazekas et al., 2020; Havron et al., 2021, 2019).

In addition to modeling, there is empirical evidence to suggest a role of prediction error as a consequence of surprisal in language learning, specifically in the development of syntactic representations. Encountering an infrequent structure (which has high surprisal) leads to stronger structural priming of that structure compared to encountering a frequent one (Bernolet & Hartsuiker, 2010; Jaeger & Snider, 2013; Kaan & Chun, 2018; Kaschak et al., 2006), a phenomenon usually referred to as "inverse frequency priming." Inverse frequency priming effects have been shown to last beyond immediate priming, leading to adaptation in L1 in both adults and children (Fazekas et al., 2020; Jaeger & Snider, 2013). Fazekas et al. (2020) investigated adaptation to the English dative alternation (direct object vs. prepositional dative construction) in an empirical study with both adults and children. They found that exposing participants to surprising dative sentences (using verbs rarely associated with the dative structure) made participants more likely to use the dative structure in a post-test.

Empirical evidence for error-based learning in L2 acquisition

Alongside L1 acquisition research, the evidence reviewed in the previous section has led to increasing interest in the role that prediction error may play in second language (L2) acquisition too (Bovolenta & Marsden, 2021b; Kaan & Grüter, 2021; Phillips & Ehrenhofer, 2015). Crucially, inverse probability priming and adaptation effects have also been observed in L2 speakers (Kaan & Chun, 2018; Montero-Melis & Jaeger, 2020) suggesting that error-based learning mechanisms may be active during L2 acquisition. Priming effects in L2 learners can be affected by the statistical

distribution of relevant structures in the learners' L1, especially at lower proficiency levels (Jackson & Ruf, 2017; Montero-Melis & Jaeger, 2020). In Montero-Melis & Jaeger (2020), L2 Spanish (L1 Swedish) speakers were exposed to descriptions of motion events that varied in how they were encoded (by path or manner). For low-proficiency speakers, adaptation was strongest for encoding that was rarer in their L1 Swedish, but as proficiency increased, learners progressively aligned with L1 Spanish speakers, that is with stronger adaptation to the type of encoding that is rarer in Spanish than in Swedish. Therefore, it seems that low-proficiency learners can exhibit inverse frequency priming based on the statistical distribution of the relevant structure in their L1 and gradually become sensitive to L2 statistics as their proficiency increases. However, while these findings provide evidence of a shift in the strength of established L2 representations, they do not provide direct evidence for a role of prediction error in the development of *new* syntactic representations. To our knowledge, no study has investigated inverse frequency priming and adaptation effects at the earliest stages of L2 acquisition.

Evidence from artificial language learning studies suggests that *direct* structural priming effects can operate at the very earliest stages of L2 acquisition in adults: in Weber et al. (2019), participants who were exposed to a novel artificial language began exhibiting repetition priming for syntactic structures from the second day of exposure, measured by faster read-aloud times and improved structural comprehension on a picture matching task. Therefore, it is of theoretical interest to investigate whether inverse probability effects could a) lead to enhanced priming at the earliest stages of L2 acquisition and b) have lasting effects on newly developed representations, promoting the establishment of structural knowledge. To our knowledge, this question has not been investigated before. If we observe that inverse probability priming and adaptation can affect the development of new structural representations, it could suggest that error-based learning mechanisms can operate at the initial stages of L2 learning in adults.

Previous empirical studies on priming, including inverse frequency priming, have usually relied on the distribution statistics of competing syntactic structures, such as the alternation between the propositional dative and direct object dative constructions in English (Fazekas et al., 2020; Jaeger & Snider, 2013; Kaschak et al., 2011). However, for ab initio learners, one might ask what the source of prediction error would be. On the one hand, evidence suggests that priming effects in low-proficiency L2 learners are affected by the statistics of related constructions in their L1 (Montero-Melis & Jaeger, 2020; Weber et al., 2019). On the other hand, the distribution of the L2 input can inform learners' expectations even at the earliest stages of learning. For instance, artificial language learning research on the acquisition of verb selectional restrictions has shown that the presence of a class of alternating verbs (i.e., verbs that can occur with different syntactic structures) in an artificial language can affect the acquisition of other verbs, generating weaker selectional restrictions for non-alternating verbs learned in alternating context relative to those learned in a fully non-alternating one (Wonnacott et al., 2008). Relatedly, formal accounts of generalization in the development of linguistic rules, including syntactic alternation (Yang & Montrul, 2017), suggest that the extent to which learners generalize new rules depends on the ratio between the total number of items in a category (e.g., verbs), and the number of instances from that category

that do and do not conform to the rule (e.g., verbs that can alternate between competing syntactic structures versus those that cannot). Until a threshold for generalizing a rule is crossed, learning remains item-specific. Therefore, the distribution of a rule in the input can shape rule learning to be item-specific, creating a potential source of prediction error. In the current investigation, we used the alternation between the active and passive structure in an artificial language as a case study. We manipulated surprisal values for verbs in specific syntactic contexts by only exposing participants to non-alternating verbs during initial learning, which would generate strong expectations for verbs to be structure-specific—providing the opportunity for prediction error when these expectations would later be violated.

The current study

The aim of this study was to test whether manipulating input surprisal could aid the acquisition of new L2 syntactic structures. The specific mechanism we investigated was inverse frequency priming and adaptation, which we assumed to be an instance of implicit error-based learning (Chang et al., 2006). We hypothesized that if inverse frequency effects can occur at the earliest stages of developing L2 syntactic representations, we should see immediate and delayed priming effects for high-surprisal verb-structures as manifested by higher accuracy in structural comprehension (Weber et al., 2019), as well as grammaticality judgments. To address our research question, we conducted a pre-registered study, in which participants learned an artificial language over the course of three days.

The language and training paradigm we used were built on a previous language learning study, which investigated the effect of prediction error at the event level (Bovolenta & Marsden, 2021a). In that study, participants learned an artificial language with an active and a passive structure (Yorwegian). Learning took place in a cross-situational learning paradigm where participants heard sentences and had to select their correct interpretation from two pictures presented on screen. Cross-situational learning is uninstructed and exposes learners to the language under conditions of uncertainty, in a way that reflects, to some extent, naturalistic language learning (Rebuschat et al., 2021; Walker et al., 2020; Yu & Smith, 2007). Bovolenta & Marsden (2021a) aimed to generate prediction and prediction error by manipulating feedback to participants' answers, whereby the feedback either aligned with or violated expectations. In the current experiment, we changed Bovolenta & Marsden's paradigm to study the effect of verb surprisal on priming by manipulating the statistical distribution of verbs in the language (instead of manipulating the syntactic structure used in feedback).

Training on the first day established expectations for specific co-occurrence patterns between individual verbs and structures, which were then violated on the second day for the surprisal group, but not for the control group. Participants were then tested on their knowledge of the Yorwegian active and passive structures using old (already encountered) as well as new (not previously encountered) verbs to test for generalization.

Research questions and hypotheses

Our main research question was whether higher verb surprisal would lead to inverse probability priming and adaptation for newly encountered structures. We hypothesized that high-surprisal input would lead to inverse frequency priming and adaptation even at the very earliest stages of language acquisition, promoting the development of new structural representations. If higher surprisal led to priming, we would expect to see an immediate (priming) effect as well as a delayed one (adaptation). We tested for priming effects on acquisition with two kinds of auditory tests: structural comprehension (both immediate [day two] and delayed [day three]) and grammaticality judgments (delayed only).

With regard to grammaticality judgments, we also hypothesized that encountering verbs in unexpected syntactic contexts may make the surprisal group more likely to revise their expectations and accept verbs in alternative structures, compared to the control group. Therefore, we expected the surprisal group to be more accepting of verb-mismatched items (e.g., formerly—Day 1—active verbs presented on Day 2 in passive structures) in the auditory grammaticality judgment task relative to the control group.

Data availability

All materials, data, and analysis code for the experiments in this article can be found at https://doi.org/10.17605/OSF.IO/EU4AV and on the IRIS database (https://www.iris-database.org/).

Method

The predictions, sampling plan, and statistical analysis for this study were pre-registered online (https://doi.org/10.17605/OSF.IO/Q9KRZ).

Power analysis

To calculate sample size, we ran a power analysis based on the findings of a previous study carried out using the same paradigm, though with different statistical distributions on Day 1 (Appendix S1). That study had shown group differences in a test of structural comprehension at the end (Day 3), with higher accuracy on passive structures for the surprisal group, but these differences were not statistically significant. We calculated Bayes' factors for the difference between means in this structural comprehension test using a Bayes' factor online calculator (Dienes, n.d., 2014). The results showed that the observed difference had a Bayes' factor of 1 (inconclusive), meaning that it did not provide strong evidence either in favor of or against our hypothesis. Given the trends we observed, we considered whether the manipulation we used may not have been sufficiently strong: evidence suggests that adaptation effects can be quite subtle and that studies examining these effects require large numbers of participants in order to reach acceptable statistical power (Prasad & Linzen, 2021).

The R script for the power analysis is available from the OSF repository for this study (https://doi.org/10.17605/OSF.IO/EU4AV). We simulated an average Surprisal - Control difference of 8% on passive sentences and -2% on active

sentences. We tested for an interaction between group and structure using a GLMER with random intercepts for subjects and items. The results showed that increasing power by using a larger sample size would be impractical: a sample size of 144 would be required to achieve .80 power. Therefore, we opted instead to increase the number of testing items (k). Our simulation showed that if we tripled the number of items used in the structural comprehension tests, a sample of 84 participants would achieve .97 power to observe a significant interaction of the size observed in our preliminary experiment.

Participants

84 native speakers of English (68 females, $M_{AGE} = 33$, $SD = 6.31$, range 18–45) were recruited via the online research platform Prolific (https://www.prolific.co/) and completed the study over the course of three consecutive days, receiving a compensation of £12. The study was given ethics approval by the Ethics Committee in the Department of York at the University of York. Participants all reported living in the United Kingdom at the time of taking part in the study, and all had English as their first and home language. All had to be 18 or over. 13 out 84 reported being university students. None of the participants reported having any knowledge of Scandinavian languages. On the first day of the study, participants were randomly assigned to either the surprisal or control group.

Stimuli

Participants were trained in an artificial language (Yorwegian), consisting of four nouns (*glim, blom, prag, meeb*—man, woman, boy, girl), twelve verbs (*flug-, loom-, gram-, pod-, zal-, shen-, norg-, klig-, jeel-, lemb-, gond-,* and *vang-*—to call, chase, greet, interview, pay, photograph, scare, threaten, dismiss, serve, kick, tease), one determiner (*lu* - the) and one preposition (*ka* - by), following the stimuli used by Bovolenta & Marsden (2021a). The specific word-meaning pairs, within the noun and verb categories, were randomly assigned for every participant. All sentences were SVO, but there were two possible syntactic structures, differentiated by verbal inflection and use of the preposition *ka*. These were the active structure (e.g., *Lu meeb flugat lu prag,* meaning, for example, "The girl greets the boy") and the passive (e.g., *Lu prag fluges ka lu meeb,* "The boy is greeted by the girl"). The two structures are modeled on the active and passive structure found in Norwegian (as well as other Scandinavian languages).[1] The rationale for using these structures is that while the active/passive alternation is familiar to L1 English speakers, the Norwegian passive structure is formed in a different way to the English one (by verb inflection instead of a BE auxiliary + participle). This choice ensured that the passive structure in the study could not be learned simply by directly transferring the L1 English structure wholesale.

Sentence stimuli were accompanied by the set of 288 black and white photographs used by Bovolenta and Marsden (2021a), which those authors had adapted from materials created by Segaert and colleagues (Menenti et al., 2011; Segaert et al., 2012). The photographs depicted transitive scenes involving the twelve verbs and four nouns of Yorwegian. Each action (e.g., *call*) was played out in twelve different agent-patient combinations (*man call woman, woman call man, man call*
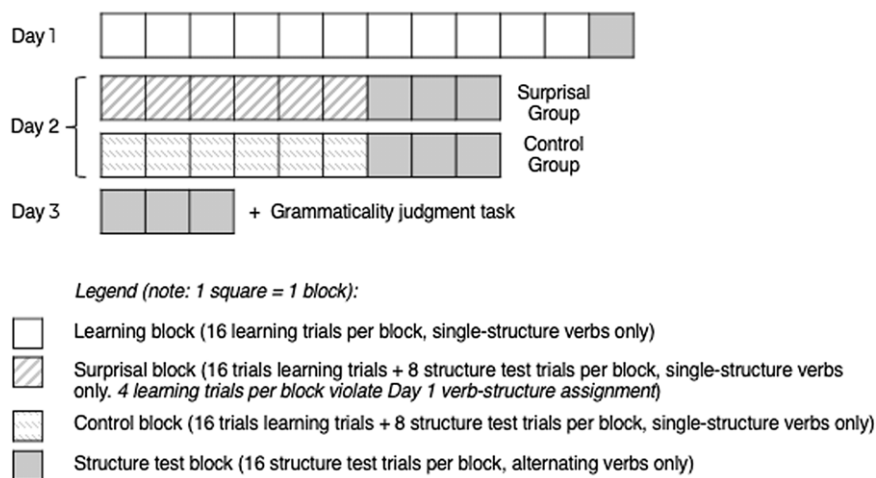
Figure 1. Summary of experimental procedure.

boy, etc.), and there were two versions of each combination, enacted by different pairs of actors.

In the learning blocks on Day 1 and 2 (including the target structure test trials on Day 2), participants were exposed to eight verbs. These verbs could only occur with one of the structures (single-structure verbs): four verbs always appeared in the active, the other four always in the passive. Four more verbs were then introduced in the structure testing blocks at the end of Day 2 and 3, and in the grammaticality judgment task. These latter four verbs could occur equally frequently with either structure (alternating verbs). Because participants had not been exposed to them during training, the four alternating verbs used in the tests served as a test of how well participants could generalize their structural knowledge to new instances.

Procedure

Participants took part in the study online over the course of three consecutive days. The average total duration of the study was ∼75 min, with each of the three sessions taking approximately 25 min. On Day 1, Day 2, and Day 3, participants performed an auditory cross-situational learning task (Figure 1), which included both learning trials and structural comprehension test trials. On Day 3, participants also did an auditory grammaticality judgment task and filled in a debriefing questionnaire. All tasks were created using JavaScript library PsychoJS, based on PsychoPy (Peirce et al., 2019). All experimental scripts were hosted and run online through the platform Pavlovia (https://pavlovia.org/). Surveys (to gather data on participants' language background and awareness of Yorwegian rules at the end of the experiment) were administered using Qualtrics (www.qualtrics.com).

Cross-situational learning task

Participants received no explicit instruction on either the grammar rules or vocabulary of Yorwegian. Participants heard individual sentences in Yorwegian,

Figure 2. Example of a learning trial and structure test trial used in the cross-situational learning task. Participants hear a sentence (written version here for display only) and must choose the correct picture by pressing the arrow keys on their keyboard.

while two pictures (a target picture and a distractor picture) appeared on screen side by side. Their task was to select the picture that corresponded to the sentence they just heard (the target) by pressing the left or right arrow on their keyboard. Thus initially, responses would be based on guessing, but participants would then gradually gather more evidence to allow them to make more informed choices. There were two types of trials: *learning trials* and *structure test trials* (Figure 2).

In *learning trials*, the agent, patient, and verb depicted in the distractor picture were selected by the software at random, with the only constraint being that the distractor verb could not be the same as the target verb (to avoid the possibility of participants seeing two pictures depicting the same scene, only enacted by different actors). These trials were designed to expose participants to the language, including co-occurrence patterns between verbs and structures, in a semi-naturalistic way.

In *structure test* trials, the same nouns and verb were depicted in both target and distractor picture, but with reversed agent and patient roles (e.g., if the target picture depicted *The girl interviews the man*, the distractor would depict *The man interviews the girl*). These trials tested whether participants could assign the correct interpretation to each structure (active and passive). The position of agent and patient characters inside the pictures (left/right) was randomized, as was the position of target and distractor pictures on screen (left/right).

Design of trials, blocks, and sessions in the cross-sectional learning task. On Day 1, all participants followed the exact same protocol, with 176 learning trials (11 blocks of 16), evenly split between active and passive sentences. The training items were created from a set of eight "single-structure" verbs, which only ever occurred in one of the two structures (four in the active, four in the passive; Table 1). Learning trials were followed by a structure test block also using single-structure verbs (16 items). At this stage, participants were not given any feedback on their answers, in either the learning or structure test trials.

On Day 2, participants did 96 learning trials (six blocks of 16). Eight of the trials in each block of 16 were followed by feedback (after the participants made their choice, the correct picture was again displayed in the center of the screen, and the sentence was played again) and then by a structure test trial. Half of the trials that were followed by feedback (i.e., four per block) were normal learning trials, and each

Table 1. Distribution of verbs used in the cross-situational learning task

	Single-structure verbs	Alternating verbs
Day 1	8 (4 passive, 4 active)	0
Day 2	Control group: same as Day 1, Surprisal group: ¼ of trials invert verb-structure assignment	4 (in structural test blocks only)
Day 3	–	4 (same as Day 2)

structure test trial that followed them simply tested participants' structural knowledge ("neutral structure test" trials). The other half of the learning trials with feedback (i.e., four per block) was where the surprisal manipulation was implemented: for the surprisal group, these trials used a single-structure verb with the opposite structure (e.g., a formerly [Day 1] "active-only" verb would now be presented in a passive sentence). The corresponding trials in the control group used the appropriately consistent structure (e.g., a formerly [Day 1] "active-only" verb was presented in an active sentence, consistent with the Day 1 learning phase). The structure test trials that followed these manipulated trials ("critical structure test" trials) were aimed at testing immediate priming effects. There were four neutral and four critical structure test trials in each block, for a total 24 neutral and 24 critical trials over the course of the Day 2 session. After the learning phase, participants did a structure comprehension test using novel alternating verbs, which consisted of 48 items (split into three blocks of 16).

On Day 3, a second structure comprehension test with the same alternating verbs as used on Day 2 was administered, also of 48 items over three blocks of 16.

Grammaticality judgment task

After the cross-situational learning task on Day 3, participants did an auditory grammaticality judgment task (a widely used technique—see Plonsky et al., 2020) with Yorwegian sentences. They were instructed to listen to each sentence and indicate whether it was a correct sentence in the language they had been learning. After each sentence was played, the words CORRECT and INCORRECT appeared side by side on screen, and participants had to press either the left or right arrow on their keyboard to give a response. Responses were untimed and the next sentence was heard only after participants gave a response. Participants heard a total of 96 sentences, of which 48 were grammatical and 48 ungrammatical. Sentences were evenly distributed between verb types (alternating and single-structure) and structures (active and passive). Ungrammatical active sentences contained the active verbal inflection incorrectly followed by the preposition *ka*, while ungrammatical passive sentences contained the passive verbal inflection but no preposition (Table 2). While this operationalization of grammaticality and ungrammaticality was arbitrary (as, for example, an active verbal inflection followed by the preposition *ka* could be labeled as "ungrammatical passive"), the critical distinction was that the structures were "ungrammatical"—albeit in different ways—relative to the language that participants had been exposed to.

Table 2. Sample items from the grammaticality judgment tasks

	Grammatical	Ungrammatical
Active inflection	*Lu meeb flug**at** lu prag* "The girl greets the boy"	*Lu meeb flug**at ka** lu prag* "The girl greets by the boy"
Passive inflection	*Lu prag flug**es ka** lu meeb* "The boy is greeted by the girl"	*Lu prag flug**es** ⊘ lu meeb* "The boy is greeted ⊘ the girl"

Debriefing questionnaire

At the end of Day 3, participants filled in a language background and debriefing questionnaire. The first part of the questionnaire included questions on the participants' educational and language background, including the amount of formal grammar instruction received in the L1, whether participants could speak any foreign languages, and the amount of instruction received in any foreign languages spoken. The second part included specific questions on the experiment itself, aimed at probing participants' awareness of the structures and of the functional distinction between them ("Did you notice that a new type of sentence was introduced on Day 2 (yesterday's session)?", and if Yes, "What were the two types of sentences you learned, and what do you think the difference was between them?").

Statistical analysis

We analyzed data with mixed-effects modeling implemented in R version 4.0.3 (R Core Team, 2021). Accuracy data[2] from structure tests and endorsement data from the grammaticality judgment task were analyzed with generalized linear mixed-effect models (GLMER) for binomial data, using the *lme4* package (Bates et al., 2015).

We used dummy coding for all categorical variables. For fixed effects, the model for structure tests included group (control: 0, surprise: 1) and structure (passive: 0, active: 1) as fixed predictors. The models for the grammaticality judgment task included group, grammaticality (grammatical, ungrammatical), and verb inflection (active, passive)[3] as predictors. Target structure tests contained only alternating verbs, whereas the grammaticality judgment task contained both single-structure and alternating verbs. Therefore, endorsement data from the grammaticality judgment task were analyzed in two separate GLMER models: The first model was on alternating verb trials only (ensuring that results could be compared with data from the structure tests, which used alternating verbs only), with group, grammaticality, and verb inflection (active vs. passive) as predictors. The second model included all trials, with verb-structure (mis)match (i.e., whether or not the verb had been used with that inflection during Day 1 training) added as predictor. We also computed *d'* scores for the grammaticality judgment task (the difference between correctly accepted grammatical items and incorrectly accepted ungrammatical ones) as a measure of grammatical sensitivity independent of individual bias. We analyzed *d'* scores in a multiple linear regression with group and verb inflection as predictors.

Table 3. Descriptive summary of main data from debriefing questionnaire

	Has L2 knowledge[*]		Passive and active structure awareness	
	Yes	No	Yes	No
Control ($n = 42$)	26	16	11	31
Surprisal ($n = 42$)	26	16	10	32

[*]At any level and regardless of how the knowledge was acquired (question: "Do you have any knowledge of any languages in addition to English?").

When constructing the mixed-effects models, we used the maximal random structure supported by the model, following Barr et al. (2013). For each model, we first created a formula containing the maximal fixed effect structure and the maximal random effect structure (random intercepts by subject and item, as well as random slopes for subjects and items by each of the fixed effect predictors, and their interactions). We identified the maximal random structure that would allow the model to converge using the package *buildmer* (Voeten, 2020). We then used *buildmer* again on the resulting formula do stepwise backward model selection using likelihood-ratio tests, eliminating fixed effect predictors one by one (starting from higher-level interactions) and only retaining them if they significantly improved model fit. All models were checked for overdispersion and none of them showed signs of being overdispersed. Any post-hoc comparisons were carried out using the *emmeans* package (Lenth et al., 2021). We report the coefficients of the mixed-effects models converted to odds ratios (*OR*) to provide a measure of effect size, together with the statistical significance of the effects (*p* values), with α = .05.

In addition to the pre-registered analysis outlined above, we carried out a number of exploratory analyses, which we report together with the corresponding pre-registered analysis (specifying clearly that they are exploratory).

Results

Descriptive statistics for our participants can be found in Table 3. The groups were matched in L2 learning experience, and they did not differ in their awareness of the function of the two Yorwegian structures at the end of the study (operationalized as being able to describe the function of the structures, and/or being able to provide correct translations of sentences using the structures with novel verbs). A full summary of data from the debriefing questionnaire can be found on the OSF repository for this study.

Below, we report the results of our statistical analyses. A summary of findings from pre-registered and exploratory analyses can also be found in Table 4; full model outputs can be found in Appendix S2. Error bars in all figures represent 95% confidence intervals.

Table 4. Summary of main statistically significant effects from pre-registered and exploratory analyses

Test	Pre-registered analysis	Exploratory analyses
Day 1: Structure test block	Main effect of group (not predicted)	
Day 2: Structure test trials	No effect of group	
Day 2: Structure test block	No effect of group	
Day 3: Structure test block	No effect of group	Day 2 accuracy added as covariate: group × structure interaction
Grammaticality judgment task	Endorsement (m1): Group × grammaticality interaction Endorsement (m2): Verb-structure match* added as covariate: Group × verb match × inflection interaction d': main effect of group	

*Verb-structure match: whether verb-structure pairing follows or violates Day 1 verb-structure assignments.
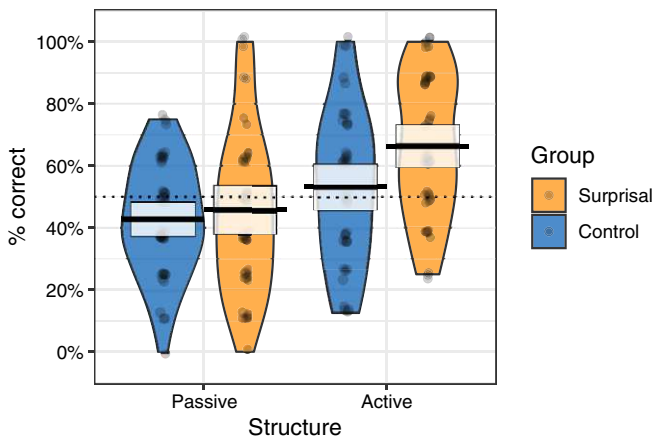
Figure 3. Average accuracy on Day 1 structure test block (k = 16).

Cross-situational learning task: Structural comprehension

Day 1: Structure test block (single-structure verbs): Baseline structural comprehension test

The structure test at the end of Day 1 took place before the surprisal manipulation was introduced, so we expected both groups to perform similarly. However, we observed significant differences between the groups as the surprisal group showed higher accuracy (Figure 3). We observed a main effect of group ($OR = 1.41$, 95% CI [1.03, 1.95], $p = .034$), as well as one of structures ($OR = 2.04$, 95% CI [1.47, 2.83], $p < .001$), due to overall higher accuracy for active sentences. We discuss possible
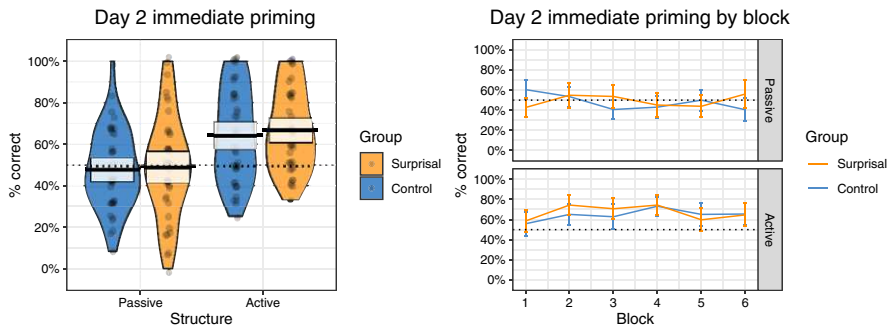
Figure 4. Mean accuracy on structure test target trials during Day 2 learning task (blocks 1–6), aggregated (left panel) and by block (right panel).

reasons for the unexpected differences between groups at baseline in the Discussion (Limitations section).

Day 2: Structure test trials during learning (single-structure verbs): Immediate priming test

If high verb surprisal increased immediate priming effects (inverse probability priming), we expected to see a main effect of group in immediate priming test trials, with the surprisal group showing higher accuracy than the control group. We entered data from all target structure test trials during learning (blocks 1–6) in a GLMER model with group and structure as predictors. We observed a main effect of structure, with overall greater accuracy for active sentences ($OR = 2.27$, 95% CI [1.62, 3.18], $p < .001$) but no effects of group, meaning that the group difference observed on Day 1 was no longer present (Figure 4). We did not, therefore, observe evidence of immediate priming, nor a visible learning effect over the course of the Day 2 learning task.

Day 2: Structure test blocks (alternating verbs): Same-day structural comprehension test

In comprehension tests following exposure, we hypothesized that if high verb surprisal contributed to adaptation to novel structures, we should see a main effect of group[4], with higher accuracy for the surprisal group relative to control. In the structure test blocks at the end of Day 2 (blocks 7–9), we observed an effect of structure, with higher accuracy for active sentences ($OR = 5.61$, 95% CI [3.30, 9.54], $p < .001$) but no significant main effects of group or interactions between group and structure (Figure 5).

Day 3: Structure test blocks (alternating verbs): Delayed structural comprehension test

In the delayed comprehension test on Day 3, as in the Day 2 comprehension test, we expected to see a main effect of group, with higher accuracy for the surprisal group relative to control. Although there was a visible trend towards an interaction
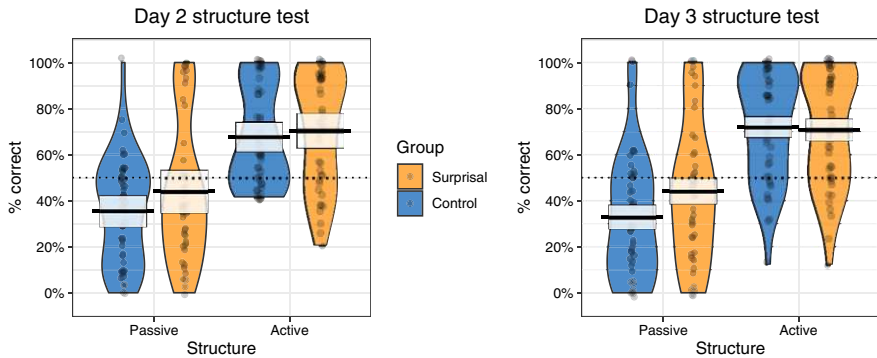
Figure 5. Mean accuracy on Day 2 structure test blocks (blocks 7–9 of Day 2 task) and Day 3 structure test blocks.

between group and structure (Figure 5), it was not statistically significant in the pre-registered analysis, which returned only a main effect of structure ($OR = 7.70$, 95% CI [4.08, 14.54], $p < .001$).

Given the variability between groups observed on Day 1, we ran an exploratory analysis to get a more sensitive measure of the change in participants' knowledge from Day 2 to Day 3, adding accuracy on Day 2 test trials as a covariate. The rationale for using these trials as a baseline measure is that they provide the earliest picture of participants' structural knowledge after the chance for overnight consolidation, just prior to further exposure and the manipulation on Day 2, and it had a higher number of items (24 instead of 16) relative to the Day 1 structure test block. The lack of differences between groups in the structure test trials on Day 2 (Figure 4) suggests that they were not affected by the group manipulation, also rendering them suitable as a baseline measure.

When adding accuracy on Day 2 structure test trials as a covariate to the model, we observed significant interactions between group and structure ($OR = 0.28$, 95% CI [0.09, 0.87], $p = .028$) and between group and Day 2 accuracy ($OR = 2.04$, 95% CI [1.18, 3.52], $p = .010$)[5]. Post hoc comparisons showed that the interaction between group and structure was due to a significant difference between groups on the passive items ($OR = 2.63$, 95% CI [1.21, 5.69], $p = .014$) but not on the active items. Therefore, we observed a significant effect of the surprisal manipulation on comprehension, which affected passive items but not active items. Post hoc tests on the interaction between group and Day 2 accuracy showed that the effect of Day 2 accuracy on Day 3 accuracy was significant for both groups (surprisal: $\beta = 1.51$, 95% CI [1.12, 1.90], $p < .001$; Control: $\beta = 0.80$, 95% CI [0.41, 1.18], $p < .001$), but the effect was smaller in the Control than in the surprisal group ($\beta = -0.71$, 95% CI [−1.26, −0.17], $p = .010$).

Aural grammaticality judgment task: Structural knowledge and verb selectional restrictions

If high verb surprisal contributed to adaptation to the novel structures, we expected the surprisal group to show better structural knowledge relative to control. In the
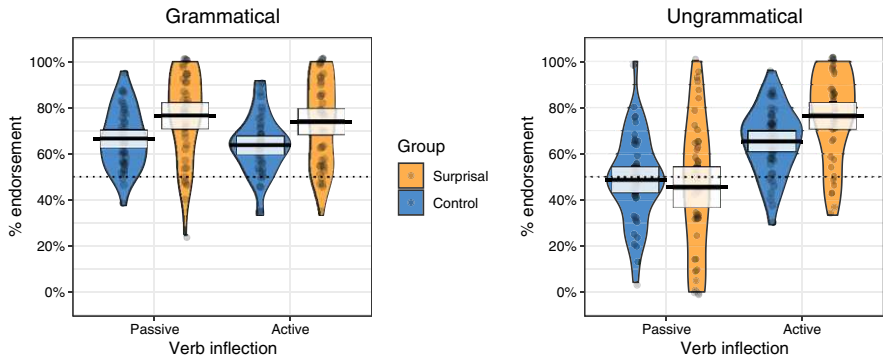
Figure 6. Average endorsement in the grammaticality judgment task (all items), by sentence grammaticality.

grammaticality judgment task, we therefore expected to see a group × grammaticality interaction: the surprisal group should be more likely to endorse grammatical sentences as grammatical, and less likely to endorse ungrammatical ones as grammatical relative to control. Analyzing endorsement of items with alternating verbs (i.e., the four alternating verbs that were introduced in Day 2)[6], we observed significant two-way interactions between group and verb inflection ($OR = 1.51$, 95% CI [1.08, 2.21], $p = .017$), between grammaticality and group ($OR = 1.77$, 95% CI [1.10, 2.87], $p = .02$), and between grammaticality and verb inflection ($OR = 0.30$, 95% CI [0.21, 0.42], $p < .001$)[7]. Overall, the surprisal group showed higher endorsement of all item types compared to control, apart from for ungrammatical passive sentences, i.e., sentences with the passive verb inflection but n *ka* marker (Figure 6). This means that participants in the surprisal group were more accurate in accepting all grammatical sentences, but they were also less accurate than control in rejecting ungrammatical active ones.

We analyzed *d'* scores (Figure 7) to assess sensitivity to grammaticality. This analysis included all items (both the four alternating and the eight structure-specific verbs), as per the pre-registration. When entering the scores in a linear regression with group and verb inflection as predictors, we observed a significant effect of group ($b = 0.43$, 95% CI [0.06, 0.80], $p = .023$), due to higher *d'* scores among the surprisal group, as well as a main effect of verb inflection ($b = -1.11$, 95% CI [$-1.48, -0.74$], $p < .001$), due to higher discrimination accuracy for sentences in the passive inflection. The results thus show a significant effect of the surprisal manipulation on the development of structural knowledge[8].

We then analyzed endorsement for structure-specific items, to test our secondary hypothesis that the surprisal group would be more accepting of verb-mismatched items relative to the control group, as they would have adapted to be more accepting of verbs alternating between either structure to a greater extent than control (Figure 8). Following the pre-registered analysis, we added verb-structure match to the model together with group, grammaticality, and verb inflection. We found a three-way interaction between group, verb-structure match, and inflection ($OR = 0.25$, 95% CI [0.15, 0.40], $p < .001$). Post hoc comparisons showed that participants in the surprisal group were more likely than those in the control group
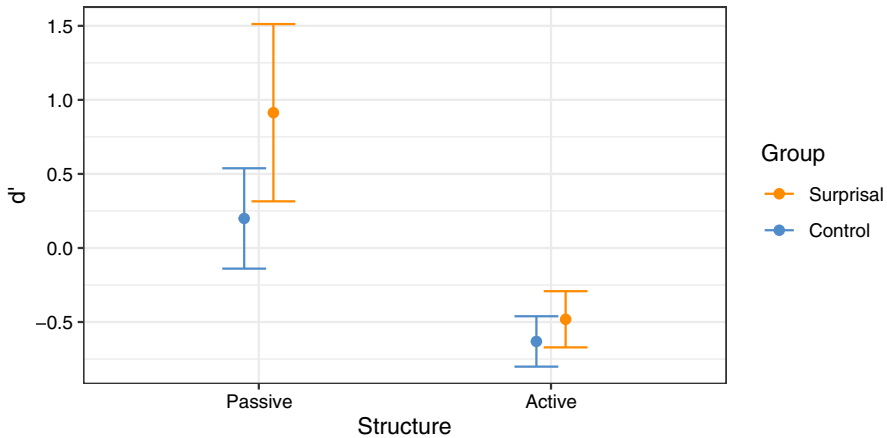
Figure 7. Mean d' scores in grammaticality judgment task by group and verb inflection.
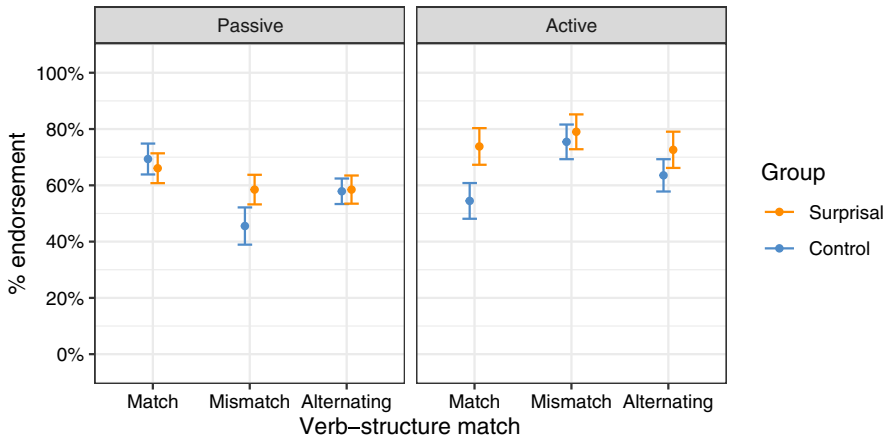

Figure 8. Breakdown of endorsement rates in grammaticality judgment task based on match between verb type and inflection (structure) used, aggregated across grammatical and ungrammatical items. Single-structure verbs are divided into "Match" (appropriate verb for that structure) and "Mismatch" (verb that had been used with the opposite structure during learning phase on Day 1).

to accept verb-mismatched items using the passive inflection ($OR = 1.88$, 95% CI [1.31, 2.68], $p < .001$) (i.e., those verbs that had only been encountered with the active structure during training, with the exception of surprisal trials), in line with our hypothesis. Participants in the surprisal group were also more likely than control to endorse verb-matched items with the active inflection, which was not predicted: $OR = 2.62$, 95% CI [1.86, 3.77], $p < .001$. Results for the passive structure suggest that experiencing prediction error during learning led participants to revise their expectations. Again, this was limited to the passive structure only, mirroring findings from the Day 3 structural comprehension test and d' scores.

Discussion

We had hypothesized that being exposed to high-surprisal input would generate prediction error and lead to inverse frequency priming and adaptation effects in the surprisal group relative to control. Specifically, we expected the surprisal group to show higher accuracy in both immediate and delayed tests of structural comprehension, and in a delayed grammaticality judgment task.

Our results provide partial support for our hypothesis. We did not observe any immediate priming effects, nor any effects in a structural comprehension test immediately following training on Day 2. On Day 3, we observed significant effects of surprisal on structural comprehension, although these only emerged in an exploratory analysis with Day 2 accuracy added as covariate (and not in the pre-registered analysis or an alternative analysis with Day 1 accuracy as covariate, possibly due to the unexpected between-group differences found on Day 1).

By contrast, findings from the grammaticality judgment task were more robust. We observed significant effects of surprisal on endorsement and accuracy (d') in grammaticality judgments (which were replicated when controlling for Day 1 and Day 2 accuracy) and on the strength of verb selectional restrictions. These results indicate that the surprisal condition had promoted knowledge of grammatical structure form (i.e., the combinations of noun order, verb inflection, and preposition use characterizing the active and passive structure) and had also led learners to update their expectations for verb-structure co-occurrences. The results from structural comprehension tests and grammaticality judgments suggest that experiencing high-surprisal input increased adaptation to newly encountered structures, promoting the establishment and development of structural representations. Unexpectedly, the effects—in both structural comprehension tests and grammaticality judgment tasks—were only observed on the passive structure, even though the manipulation was applied to both structures. We discuss possible interpretations for these findings below, as well as potential limitations of the current study.

Effect of surprisal on passive structures only

In this study, we observed an effect of verb surprisal, but only on the passive structure—even though both structures underwent the surprisal manipulation. This finding was not predicted by our hypothesis. One possibility is that this finding may simply be due to a ceiling effect for active sentences. We can speculate that active sentences, being by far the more frequent structure in the participants' native language (English), would also be easier to acquire than the passive. The Yorwegian active structure is also constructed in the same way as the English one (unlike the passive), yielding a potential L1 transfer advantage. Additionally, a preference for the active structure is not only a feature of English, but has been attested cross-linguistically in children (Estevan, 1985; Jakubowicz & Seguí, 1980; Maratsos et al., 1985). Finally, the entities that served as subjects and objects in our study were all animate and therefore likely to be interpreted as agents during sentence processing (Hare et al., 2009; Kim & Osterhout, 2005). Therefore, participants may have defaulted to an active interpretation, leading to high accuracy for active sentences

and generally low accuracy for passive sentences (while accuracy was higher in the surprisal group, it should be noted that both groups were below chance level in their comprehension of passives).

However, data from grammaticality judgments on Day 3 suggest a more complex picture: while accuracy in comprehension tests was always significantly higher for active sentences, accuracy in the grammaticality judgment task (*d'* scores) was significantly *lower* for active sentences, in both groups. Participants in both groups were equally likely to endorse active sentences regardless of their grammaticality, suggesting that they uncritically tended to accept items that contained the active verbal inflection (*-at*)[9]. The effect of high-surprisal input on verb selectional restrictions, too, only seemed to apply to endorsement of passive items. Relative to the control group, participants in the surprisal group became more accepting of passive sentences containing active-only verbs ("mismatch" items in the passive condition), regardless of grammaticality, but they did not become more accepting of active sentences with passive-only verbs ("mismatch" items in the active condition). This suggests that being exposed to mismatched verbs during the surprisal phase had led participants to revise their expectations for the passive structure (becoming more accepting of previously unattested verbs appearing in this structure), but not for the active structure.

Taken together, these data suggest a striking possibility: that participants did not develop a distinct structural representation for the Yorwegian active structure, due to its closeness to the default structure in their L1. While the passive structure was different from the English passive (most notably, due to the lack of BE auxiliary), the active structure could be mapped directly onto the English active structure. Therefore, it is possible that in comprehension tests, participants simply defaulted to an active interpretation (assigning subject role to the first noun, and object role to the second noun), resulting in high accuracy for active sentences and generally low accuracy in passive ones. But in grammaticality judgment tasks, they showed no sensitivity to morphosyntactic violation in active sentences, due to missing structural representations. For the same reason, encountering active sentences with passive-only verbs did not seem to elicit prediction error on Day 2 in the surprisal group (and consequently, no revision of verb selectional restrictions was observed).

A distinct but related possibility is that the presence of the active structure in the L1 led participants to generalize it, despite limited input. If participants saw the Yorwegian active as an instance of active (similar to their L1), then they would likely base their interpretation of the structure on distributional statistics from their L1, as has been observed in previous studies on adaptation in L2 speakers (Jackson & Ruf, 2017; Montero-Melis & Jaeger, 2020). This hypothesis is compatible with research on the acquisition of dative alternation in English, which follows different trajectories in L1 and L2 learners (Conwell & Demuth, 2007). Although double object datives are learned sooner in L1 acquisition, prepositional datives are acquired earlier by L2 learners. Although there appears to be a general preference for prepositional object datives overall among L2 learners, some evidence also suggests that a higher prevalence (proportional frequency) of prepositional datives in the learners' L1 could contribute to earlier acquisition of the same structure in the L2 (Agirre, 2015; Hawkins, 1987). Similarly, if participants in our study relied on the statistical distribution of the active structure in English, where the structure is highly

productive, they may have been more likely to generalize the Yorwegian active structure to new verbs too, even after limited exposure. By contrast, because no English version of the Yorwegian passive exists, the Yorwegian passive could only be acquired via item-specific learning, which would be determined by its distribution in Yorwegian. Therefore, participants may have developed stronger verb selectional restrictions for the Yorwegian passive structure than for the active one, potentially experiencing greater prediction error when these restrictions would be violated.

This explanation is compatible with theoretical accounts of the acquisition and generalization of syntactic rules. According to the Sufficiency principle (Yang & Montrul, 2017), a rule applying to a syntactic category becomes productive (i.e., there is a shift from item-based learning to generalization to the whole category) when the number of items following that rule passes a mathematically defined threshold (the difference between total occurrences of the category and the natural logarithm of the same value). In our case, the number of items (i.e., individual verbs) observed with the Yorwegian passive structure would not be sufficient for participants to *generalize* the rule (i.e., to generalize the Yorwegian passive structure to new verbs). By contrast, if participants perceived Yorwegian active sentences as instances of the active structure which they were already familiar with from their L1 English, then the number of items they had witnessed with that structure would comprise not only Yorwegian active verbs, but all English verbs they had ever encountered in the active form—a sufficient number of items to generalize the Yorwegian active structure. Under this interpretation, learners would have acquired the intended verb selectional restrictions only for the passive structure, generating prediction error when these were violated, and consequently error-based learning in the surprisal group that was restricted to the passive structure.

Lack of immediate priming effects

The other unexpected finding in our study was the fact that we did not observe any immediate effects of the surprisal manipulation, and yet we observed delayed effects. We had hypothesized that, if an error-based learning mechanism such as that specified by the Dual-Path model (Chang et al., 2006) was driving learning, we should see both immediate (priming) and delayed (adaptation) effects of prediction error. Against our predictions, however, we did not observe significantly higher accuracy on the structure test trials immediately following surprising trials, suggesting that the manipulation did not produce any immediate priming effects.

On the one hand, our results are compatible with previous findings from other studies. In their study on adaptation to alternative dative constructions (prepositional vs. direct object dative), Fazekas et al. (2020) observed adaptation following exposure to low-frequency verb-structure pairs, but no immediate inverse probability priming effects. They observed a numerical trend towards priming for adults, but not for children, suggesting that well-established representations may be needed for immediate priming effects to be elicited by prediction error. Our findings, too, suggest that it is possible for participants to experience adaptation without having shown immediate priming effects.

On the other hand, the reason for the lack of immediate priming effects in our study may lie in the specific measure we chose to measure priming, which was

structural comprehension. In an artificial language learning study, Weber et al. (2019) observed direct priming in structural comprehension starting only from the third day of an artificial learning task, while priming on read-aloud times emerged earlier in the study. Therefore, we cannot rule out the possibility that immediate priming effects may have emerged had we used a different test. Future research should investigate this possibility, using different tests of priming in order to gain a better picture of inverse frequency priming effects and how they interact with the strength of existing representations, as well as the measures used to assess priming.

Finally, the lack of immediate priming effects may simply be indicative of the fact that the advantage enjoyed by the surprisal group was not due to implicit error-based learning, but to other mechanisms—a possibility we explore below.

Alternative mechanisms for the effect of surprisal

There are a number of mechanisms by which higher surprisal could have led to greater accuracy in the surprisal group, besides implicit error-based learning. While the aim of the current study was to study the effects of prediction error on the acquisition of structures, we did not directly measure prediction error (e.g., with an online methodology such as eye-tracking). Instead, we manipulated surprisal (statistical properties of the input) with the assumption that it would generate prediction error. Therefore, while our findings are at least partially compatible with an error-based learning mechanism, they could also be explained by other types of mechanism.

One possibility is that participants were not processing the verbs they saw during training as inflected forms, but rather as whole lexical items. This would be compatible with their experience of their L1 English, where forms with identical onsets but different endings can be distinct verbs (e.g., *cont-est* and *cont-rast*). Additionally, if participants always interpreted the first noun as the agent, the preposition *ka* could be interpreted as part of an active sentence, such as introducing a prepositional complement (e.g., "The boy talks to the girl"). Crucially, this would make the presence of *ka* something related to the idiosyncratic meaning of each verb, rather than bearing a systematic relationship with a particular verb ending which could occur with multiple verbs. Under this interpretation, the surprisal group would have subjectively experienced a wider range of verbs during training, rather than the same set of verbs in more syntactic contexts. This is compatible with the findings that participants in the surprisal group have higher acceptance of ungrammatical as well as grammatical active sentences, because they may simply perceive the ungrammatical forms as new verbs (new lexical items), with a new meaning. It is also compatible with the fact that they were more accepting of active mismatched verbs (which they had already encountered during Day 2 training). However, it would not explain why the effects were structure-specific: the surprisal group was more accepting of ungrammatical active sentences, but not ungrammatical passive ones; when breaking down endorsement by verb type, the surprisal group was more accepting of mismatch in passive sentences, but not in active ones. Therefore, while it is possible that participants learned the inflected forms as whole verbs (indeed, that would have been a necessity at the start of the training, before any patterns could begin to be abstracted), results also suggest that participants eventually developed

sensitivity to the fact that different systematic patterns existed in the language. We acknowledge, however, that it is possible that the surprisal group developed a sensitivity to a lexicalized string "*es+ka*" being acceptable, rather than necessarily having established a (purely) morphosyntactic structure.

A second possibility is that abstraction itself was aided by the greater range of exemplars to which the surprisal group was exposed. More precisely, participants in this group heard a wider range of verbs in each syntactic context (because they heard the single-structure verbs in both kinds of structures), compared to control. There is evidence that variability improves learning in statistical learning tasks (Bulgarelli & Weiss, 2021; Gómez, 2002). Gómez (2002) found that the acquisition (assessed by grammaticality judgments) of non-adjacent dependencies between syllables presented in an auditory statistical learning task benefitted from greater variability in the strings intervening between syllables. Given evidence that increased variability aids learning, it is possible that the surprisal group benefitted from exposure to a wider range of verbs in each syntactic context (due to hearing the violation trials, whereas the control group did not), and this could have helped them to isolate the abstract structures from individual lexical items.

Finally, it is also possible that prediction error was indeed the cause of the observed differences between groups on Day 3, but this prediction error was not due to implicit error-based learning, and so was not observable in the immediate structural comprehension test. Instead, one possible mechanism we may have observed is one-shot declarative learning, i.e., the phenomenon that novel associations are better remembered if they violate an established pattern (Brod et al., 2018; De Loof et al., 2018; Greve et al., 2017, 2019). In language acquisition, the effect of one-shot declarative learning has been investigated in the context of vocabulary learning, both in children (Gambi et al., 2021; Stahl & Feigenson, 2017) and adults (Gambi et al., 2021). While most of the evidence comes from vocabulary learning, however, we cannot discount the possibility that one-shot declarative learning may also contribute to the development of new structural knowledge, albeit indirectly. In usage-based accounts of language acquisition, structural knowledge is thought to emerge through abstraction from individual learned exemplars (Ellis et al., 2016). Therefore, a mechanism such as one-shot declarative learning, which aids the formation of individual memories of specific instances of structure, may be hypothesized to indirectly contribute to the development of abstract structural knowledge by providing bases for generalization. To test this hypothesis, future replications of this study would need to include tests of item memory for the specific sentences heard during the training phase (see one attempt of doing this in our earlier study in Appendix S1).

Another possibility is that high-surprisal input engaged learners' attention, leading to better learning. In Bovolenta and Marsden (2021a), it was hypothesized that the observed learning effects could be due to attention raising as a function of experimental design: The feedback paradigm used for surprisal participants, which involved juxtaposing active and passive structures, could have drawn their attention to the difference between structures. The present study did not involve any juxtaposition of structures, so the same explanation could not apply. However, if surprisal caused participants to experience prediction error it may still lead to global attention raising (i.e., greater attention to the task as a whole) and overall better

learning. For instance, Fitneva and Christiansen (2011, Fitneva & Christiansen, 2017) found that accidentally experiencing prediction error (by forming incorrect label-referent mappings at the start of a cross-situational vocabulary learning task) led to overall higher learning rates in adults. The important thing to note is that the effect applied to the whole vocabulary set, not only to the words that participants had initially assigned to the wrong referent. This observation would not be compatible with implicit error-based learning, but rather suggests that higher surprisal may have led to greater attention and better encoding of information overall. The same mechanisms could potentially have played a role in the present study.

It should be noted that both of these potential mechanisms—one-shot declarative learning and attention raising—are "global," in the sense that they should in principle apply to *all* of the sentences affected by the surprisal manipulation (which were both active and passive), and would consequently be expected to boost learning of both structures. Therefore, these explanations seem at odds with our finding that effects on structural knowledge (accuracy measures) seem to emerge primarily on the passive structure. However, any of the potential reasons we explored for the lack of learning effects on the *active* structure (ceiling effects, L1 transfer) could of course still apply and so partially counteract any learning advantage derived from surprisal. Thus, this could account for the asymmetrical pattern of results we observed, even in the presence of a global learning boost.

Descriptive data from the debriefing questionnaire (Table 3) show that neither group was more likely than the other to develop awareness of the distinction between active and passive. Intuitively, one might expect greater global attention to lead to greater awareness of the rules; however, that may not necessarily be the case: research on implicit language learning shows that engaging learners' attention can affect learning even in the absence of awareness, and being unable to articulate explicit rules after a short learning study does not eliminate the possibility that attentional levels were heightened during exposure (e.g., Leung & Williams, 2006; Marsden et al., 2013).

Limitations

One notable limitation of our study was the difference observed between groups in the structure comprehension test at the end of Day 1, before the experimental manipulation was introduced. This difference (higher accuracy for the surprisal group on active sentences in the structural comprehension test) was no longer visible on Day 2 and went against the pattern consistently observed elsewhere in the experiment (where the difference between groups was on passives). In addition, the effect of group observed in the pre-registered analysis for the grammaticality judgment task was replicated in exploratory analyses controlling for both Day 1 and Day 2 accuracy (while the effect on structural comprehension only emerged when controlling for Day 2 accuracy). Therefore, we think it is unlikely that the learning effects we observed—especially in the grammaticality judgment task—were due to baseline differences between groups, but can be ascribed instead to the experimental manipulation on Day 2.

Nevertheless, observing a difference between groups on Day 1 was unexpected, given our random sampling. One tentative explanation for this difference may lie in

the fact that it is more difficult to avoid attrition in online data collection, and attrition may induce self-selection bias in terms of which participants complete the entire study. We experienced attrition rates of roughly 30% and all attritors were eliminated from the final dataset analyzed. While most attrition was due to participants dropping out after the Day 1, a few dropped out after Day 2. If the surprisal condition on Day 2 was perceived as more difficult, it could have made a particular subset of "lower performing" surprisal group participants more likely to abandon the study after Day 2 (thus leaving more of the "higher performers" from Day 1 remaining in the dataset), relative to those in the control group. However, this is a highly speculative account, and it does not explain why the initial difference between groups disappeared on Day 2. Nevertheless, we highlight this potential challenge for multi-session online research.

Conclusion

Overall, our findings indicate an effect of surprisal on the development of abstract structural knowledge. Participants who were exposed to unexpected verb-structure combinations showed higher accuracy in comprehension of the passive in delayed tests and on grammaticality judgments in delayed tests. Therefore, even at the very earliest stages of L2 acquisition, encountering a structure in an unexpected context can promote the development of structural representations. The delayed effects we observed are compatible with error-based learning accounts of language acquisition. However, we only observed the effects of group on the passive structure, even though both structures had been affected by the experimental manipulation. We suggested potential reasons for the lack of an effect on the active structure, which include ceiling effects and L1 structural biases—further research will be needed to examine these potential reasons. Also contrary to our expectations, we did not observe any immediate priming effects, which would be predicted by an *implicit* error-based learning account. The lack of immediate effects could be due to the fact that such effects may depend on more mature structural representations being already established. However, it could also indicate that a different mechanism—something other than implicit error-based learning, such as a global heightened awareness in the surprisal condition—was responsible for our findings. Therefore, further research is needed to determine the precise nature of the effect generated by our experimental manipulation and shed more light on the potential role of prediction error in L2 acquisition.

Notes

1 In Norwegian, verbs in the present tense can have either an active or passive inflection. The passive structure is formed by inverting the subject and object, and inflecting the verb in the passive form (followed by a preposition meaning *by*).
2 The experimental software we used also recorded response times (which can be found in the data in the OSF repository for the study). However, we had no hypotheses concerning response times and did not pre-register an analysis for them. An exploratory analysis of response times carried out post hoc did not reveal any effects of group and is not reported.

3 For the grammaticality judgment task, we use the term 'verb inflection' instead of 'structure' as done in previous analyses to account for ungrammatical sentences (which are not technically instances of either structure, since they mix the verb inflection of one structure with the preposition usage of the other structure).

4 Data from our preliminary unpublished study (Appendix S1), on which we based our power analysis, suggested that we might expect to see an interaction between group and structure. However, at the time of designing the current study, we had no theoretical reasons for predicting such an interaction instead of a main effect of group, because the experimental manipulation was applied to both structures.

5 An alternative exploratory analysis with Day 1 accuracy as covariate only returned the main effects of structure and Day 1 accuracy. The model for this analysis can be found in the R script for on the OSF repository for this study.

6 In the analysis we pre-registered, we decided to only include items with alternating verbs in order to get a pure measure of grammatical knowledge, to avoid potential confounds from any verb bias caused by the structure-specific verbs. The R analysis code for the study also includes a version of the model including all items, which yields a three-way interaction between group, structure, and grammaticality.

7 We could not carry out exploratory analyses adding accuracy from previous days to this model because the resulting model would not converge.

8 To ensure comparability with the exploratory analysis we carried out on structural comprehension (with Day 2 accuracy added as covariate), we also ran an additional exploratory analysis of d' with Day 2 target trial accuracy added as covariate. The results were essentially the same as the pre-registered analysis, with main effects of group ($b = 0.71$, 95% CI [0.19, 1.23], $p = .007$) and verb inflection ($b = -0.96$, 95% CI [−1.50, −0.42], $p < .001$). An additional model run with Day 1 accuracy as covariate similarly returned the main effects of group ($b = 0.70$, 95% CI [0.18, 1.27], $p = .009$) and verb inflection ($b = -0.87$, 95% CI [−1.40, −0.34], $p = .001$). Both exploratory models can be found in the R script in the OSF repository for this study.

9 It should be noted that distinguishing between active and passive ungrammatical sentences based on verbal inflection (Table 2) is somewhat arbitrary: ungrammatical sentences could equally have been coded as active or passive based on whether they contained the preposition *ka*, which would have inverted the structure categories assigned to ungrammatical sentences. Therefore, the polarity of the difference between structures in the grammaticality judgment task is not essential, but what is important to note is that there is a difference in how participants process elements associated with the active and passive structure, which suggests differences in the acquisition of the two structures.

References

Agirre, A. I. (2015). The acquisition of dative alternation in English by Spanish learners. *Vigo International Journal of Applied Linguistics*, **12**, 63–90. https://revistas.webs.uvigo.es/index.php/vial/article/view/69

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bernolet, S., & Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming? *Cognition*, **114**(3), 455–461. https://doi.org/10.1016/j.cognition.2009.11.005

Bovolenta, G., & Marsden, E. (2021a). Expectation violation enhances the development of new abstract syntactic representations: Evidence from an artificial language learning study. *Language Development Research*, **1**(1), 193–243. https://doi.org/10.34842/C7T4-PZ50

Bovolenta, G., & Marsden, E. (2021b). Prediction and error-based learning in L2 processing and acquisition: A conceptual review. *Studies in Second Language Acquisition* **44**(5), 1384–1409. https://doi.org/10.1017/S0272263121000723

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, **55**, 22–31. https://doi.org/10.1016/j.learninstruc.2018.01.013

Bulgarelli, F., & Weiss, D. J. (2021). Desirable difficulties in language learning? How talker variability impacts artificial grammar learning. *Language Learning*, **71**(4), 1085–1121. https://doi.org/10.1111/lang.12464

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, **113**(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, **103**(2), 163–179. https://doi.org/10.1016/j.cognition.2006.03.003

De Loof, E., Ergo, K., Naert, L., Janssens, C., Talsma, D., Van Opstal, F., & Verguts, T. (2018). Signed reward prediction errors drive declarative learning. *PloS One*, **13**(1), e0189212. https://doi.org/10.1371/journal.pone.0189212

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, **5**, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (n.d.). *Making the most of your data with Bayes*. Retrieved January 6, 2022, from http://users.sussex.ac.uk/~dienes/inference/Bayes.htm

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-Based Approaches to Language Acquisition and Processing: Cognitive and Corpus Investigations of Construction Grammar*. (vol. 66, Supplement 1). Wiley.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Estevan, R. A. C. (1985). La evolución de la transformación pasiva en castellano: Génesis de una situación de modificación morfo-sintáctica. *Revista espanola de linguistica aplicada*, **1**, 9–25. https://dialnet.unirioja.es/servlet/articulo?codigo=1962900

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, **7**(11), 180877. https://doi.org/10.1098/rsos.180877

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more rate word learning. *Cognitive Science*, **35**(2), 367–380. https://doi.org/10.1111/j.1551-6709.2010.01156.x

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science*, **41**(1), 141–161. https://doi.org/10.1111/cogs.12322

Frank, S. L. (2013). Uncertainty Reduction as a Measure of Cognitive Load in Sentence Comprehension. *Topics in Cognitive Science*, **5**(3), 475–494. https://doi.org/10.1111/tops.12025

Frank, S. L., & Hoeks, J. (2019). The Interaction between Structure and Meaning in Sentence Comprehension: Recurrent Neural Networks and Reading Times. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 337–343. https://doi.org/10.31234/osf.io/mks5y.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word Surprisal Predicts N400 Amplitude during Reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 878–883). Sofia, Bulgaria: Association for Computational Linguistics. https://repository.ubn.ru.nl/bitstream/handle/2066/119221/119221.pdf

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP Response to the Amount of Information Conveyed by Words in Sentences. *Brain and Language*, **140**(January), 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Gambi, C., Pickering, M. J., & Rabagliati, H. (2021). Prediction error boosts retention of novel words in adults but not in children. *Cognition*, **211**, 104650. https://doi.org/10.1016/j.cognition.2021.104650

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, **13**(5), 431–436. https://doi.org/10.1111/1467-9280.00476

Goodkind, A., & Bicknell, K. (2018). Predictive Power of Word Surprisal for Reading Times Is a Linear Function of Language Model Quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18. https://www.aclweb.org/anthology/W18-0102.pdf

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, **94**, 149–165. https://doi.org/10.1016/j.jml.2016.11.001

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology. General*, **148**(2), 325–341. https://doi.org/10.1037/xge0000498

Hale, J. (2001). A probabilistic early parser as a psycholinguistic model. Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, 1–8. https://doi.org/10.3115/1073336.1073357

Hare, M., Elman, J. L., Tabaczynski, T., & McRae, K. (2009). The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension. *Cognitive Science*, **33**(4), 610–628. https://doi.org/10.1111/j.1551-6709.2009.01027.x

Havron, N., Babineau, M., Fiévet, A.-C., Carvalho, A., & Christophe, A. (2021). Syntactic prediction adaptation accounts for language processing and language learning. *Language Learning*, **71**, 1194–1221. https://doi.org/10.1111/lang.12466

Havron, N., de Carvalho, A., Fiévet, A.-C., & Christophe, A. (2019). Three- to four-year-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child Development*, **90**(1), 82–90. https://doi.org/10.1111/cdev.13113

Hawkins, R. (1987). Markedness and the acquisition of the English dative alternation by L2 speakers. *Interlanguage Studies Bulletin (Utrecht)*, **3**(1), 20–55. https://doi.org/10.1177/026765838700300104

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee & H. S. Cairns (Eds.), Language, speech, and communication. Methods for assessing children's syntax (pp. 105–124). The MIT Press. https://psycnet.apa.org/record/1997-97174-005

Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism*, **24**(5–6), 895–911. https://doi.org/10.1177/1367006920902851

Jackson, C. N., & Ruf, H. T. (2017). The priming of word order in second language German. *Applied Psycholinguistics*, **38**(2), 315–345. https://doi.org/10.1017/S0142716416000205

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, **127**(1), 57–83. https://doi.org/10.1016/j.cognition.2012.10.013

Jakubowicz, C., & Seguí, J. (1980). L'UTILISATION DES INDICES DE SURFACE DANS LA COMPRÉHENSION D'ÉNONCÉS. *Approches Du Langage: Actes Du Colloque Interdisciplinaire Tenu à Paris, Sorbonne, Le 8 Décembre 1978*, **16**, 63.

Kaan, E., & Chun, E. (2018). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition*, **21**(2), 228–242. https://doi.org/10.1017/S1366728916001231

Kaan, E., & Grüter, T. (Eds.). (2021). *Prediction in Second Language Processing and Learning*. John Benjamins. https://www.jbe-platform.com/content/books/9789027258946

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, **18**(6), 1133–1139. https://doi.org/10.3758/s13423-011-0157-y

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition*, **99**(3), B73–B82. https://doi.org/10.1016/j.cognition.2005.07.002

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, **52**(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, **31**(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., & Singmann, H. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://cran.r-project.org/package=emmeans

Leung, J., & Williams, J. N. (2006). Implicit learning of form-meaning connections. *Proceedings of the Annual Meeting of the Cognitive Science Society*, **28**(28), 465–470. https://escholarship.org/uc/item/06s2d7bm

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Maratsos, M., Fox, D. E., Becker, J. A., & Chalkley, M. A. (1985). Semantic restrictions on children's passives. *Cognition*, **19**(2), 167–191. https://doi.org/10.1016/0010-0277(85)90017-4

Marsden, E., Williams, J. N., & Liu, X. (2013). Learning novel morphology: The role of meaning and orientation of attention at initial exposure. *Studies in Second Language Acquisition*, **35**(4), 619–654. https://doi.org/10.1017/S0272263113000296

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science*, **22**(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical Surprisal as a General Predictor of Reading Time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). EACL '12. USA: Association for Computational Linguistics. https://dl.acm.org/doi/10.5555/2380816.2380866

Montero-Melis, G., & Jaeger, F. T. (2020). Changing expectations mediate adaptation in L2 production. *Bilingualism: Language and Cognition*, **23**(3), 602–617. https://doi.org/10.1017/S1366728919000506

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, **17**(2), 357–374. https://doi.org/10.1017/s0305000900013817

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, **51**(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Phillips, C., & Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism*, **5**(4), 409–453.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, **26**(4), 583–621. https://doi.org/10.1177/0267658319828413

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **47**(7), 1156. https://doi.org/10.1037/xlm0001046

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension–an FMRI study. *Cerebral Cortex*, **22**(7), 1662–1670. https://doi.org/10.1093/cercor/bhr249

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, **163**, 1–14. https://doi.org/10.1016/j.cognition.2017.02.008

Van Schijndel, M., & Linzen, T. (2018). Modeling Garden Path Effects without Explicit Hierarchical Syntax. In T. Rogers, M. Rau, X. Zhu, and C. W. Kalish (Eds.) *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2603–2608). Austin, TX: Cognitive Science Society. http://tallinzen.net/media/papers/vanschijndel_linzen_2018_cogsci.pdf

Voeten, C. C. (2020). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. https://cran.r-project.org/package=buildmer

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, **70**(52), 221–254. https://doi.org/10.1111/1467-923X.12837

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2019). Primed from the start: Syntactic priming during the first days of language learning. *Language Learning*, **69**(1), 198–221. https://doi.org/10.1111/lang.12327

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning Context-Based Statistical Models of Language with Brain Activity during Reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 233–243). https://www.aclweb.org/anthology/D14-1030.pdf

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, **56**(3), 165–209. https://doi.org/10.1016/j.cogpsych.2007.04.002

Yang, C., & Montrul, S. (2017). Learning datives: The tolerance principle in monolingual and bilingual acquisition. *Second Language Research*, **33**(1), 119–144. https://doi.org/10.1177/0267658316673686

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, **18**(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x