Modality Matters: Generalization in Second Language Learning After Production Versus Comprehension Practice

By

Elise W.M. Hopman

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 05/11/2022

Dissertation approved by the following members of the Final Oral Committee:

Maryellen C. MacDonald, Professor, Psychology, University of Wisconsin-Madison
Joseph L. Austerweil, Associate Professor, Psychology, University of Wisconsin-Madison
Jenny R. Saffran, Professor, Psychology, University of Wisconsin-Madison
Timothy T. Rogers, Professor, Psychology, University of Wisconsin-Madison
Carrie N. Jackson, Professor, German and Linguistics, Pennsylvania State University

Table of Contents

Acknowledgements

Many people contributed to the work presented here. My now graduated honors student Emily Perez's honor thesis work has been integral to all stages of developing the study presented here. My own mentor, Maryellen MacDonald is a driving force behind this study, and my entire PhD trajectory, as well. Feedback received at different times from Jenny Saffran, Tim Rogers, Joe Austerweil and Carrie Jackson contributed to all parts of this dissertation. Many research assistants contributed to the study proposed here by providing feedback on the initial ideas, designing and creating materials, implementing the experiment, piloting the study, and providing yet more feedback on my presentations, both written and oral, of this study - thank you, Teresa Turco, Ann Chapman, Sarah Wang, Sarah Engel, Gabriela Zayas-Alom, Megan McClement, Alma Reinebach, Mackenzie Ludin, Jamie Dawson, Rubiarbriana Jamison, Yiwen Wang, Charles Rojas, River Steen, Miriam Lebowitz, Liana Keivanfar. Thank you especially to Marcela Chavez, Neha Prasad, Misty Kabasa and Danielle Miller for coding the production data, and again for lending your voice to the artificial language in this experiment, Marcela. Levi Redlin, I would not have been able to program this experiment in JavaScript without your code and your help. I thank Celeste Kidd for generously sharing the visual stimuli used in this experiment. I'm grateful for the participants who volunteered their time to do my experiments – I know that my studies are longer and more challenging than many others. Finally, a thank you to all members of LCNL, as well as all other audiences who have heard me present this work and provided me with insightful questions and helpful feedback. Whereas this is the acknowledgments section for work done on this study, it is a time-honored tradition for Dutch dissertations to contain a very detailed and personal 'dankwoord' as its own chapter at the end of the dissertation. So I wrote one of those, too (see Appendix H).

Dedication

I dedicate this dissertation to my dad, the late Wim Hopman, who didn't get to see me complete my PhD, but encouraged me to be a researcher before I had even learned to read or write, and fully supported my decision to pursue a PhD overseas in the USA during his final years alive.

Opdracht

Pappa, ik draag mijn proefschrift aan jou op. Je hebt het afronden van mijn doctoraat en de verdediging van mijn proefschrift niet meer mee mogen maken, maar moedigde mijn nieuwsgierigheid en onderzoekendheid al aan voordat ik kon lezen of schrijven. Ik ben dankbaar dat je, zelfs toen dit jouw laatste levensjaren bleken te zijn, altijd volledig achter mijn beslissing om een PhD in de VS te gaan doen hebt gestaan.

Abstract

Generalization is the ability to apply regularities to novel instances, for example, correctly guessing that the plural for the novel English word ‘wug’ should be ‘wugs’. Early language learners make overgeneralization errors like ‘mouses’, applying regularities beyond their attested uses. Theories concerned with the question of how learners learn to correctly generalize regularities, without overgeneralizing, have recently been criticized for being insufficiently mechanistic. Rule learning and statistical learning theories typically do not take into account whether that generalization is happening during production (e.g. coming up with the plural for ‘wug’), or comprehension (e.g. judging whether ‘wugs’ or ‘wugga’ sounds better as a plural for ‘wug’). However, my own prior research showed that training modality affects regularity learning, with production training leading to a more accurate ability to apply regularities to learned words. Thus, modality may provide a potential path to making theories more specific and mechanistic.

In this thesis, I first reviewed relevant literature on generalization through a task modality lens. There is very little literature directly contrasting, in a balanced manner, the effects of production versus comprehension *training* on learning and generalization. However, generalization studies have used both production and comprehension *testing* to assess generalization performance. Drawing on these results, I identified several different patterns of generalization results by testing modality. I concluded that, if there are any modality differences, production training should lead to better generalization, and production tests should be more likely to elicit overgeneralization errors.

I then designed and conducted an artificial language learning experiment that contrasts production versus comprehension *training* modality between different groups of participants, and uses both comprehension and production *testing* to assess learning and (over)generalization. People learning the artificial language with production training were better at generalizing and made fewer overgeneralization errors than people learning the artificial language with comprehension training. Surprisingly, comprehension-trained people did do better than production-trained people on vocabulary learning. Finally, people made overgeneralization errors in both comprehension and production tests. I discussed consequences for existing theories as well as practical applications of my findings for second language learners.

Literature Review on How Production Impacts Generalization

Learning the grammar of a language is hard – recent evidence shows that even native speakers still improve on grammar tasks well into their twenties and thirties (Hartshorne, Tenenbaum & Pinker, 2018). Learning the grammar of a second language is even harder – for example, even native English speakers who work as translators and teachers of Spanish don't use grammatical gender as fluently as native speakers (Grüter, Lew-Williams & Fernald, 2012). So, how can we make grammar learning more effective?

We recently showed initial evidence that language production training improved grammar learning compared with language comprehension training, and posited that this benefit of production training was due to inherent processing differences between language production and language comprehension tasks (Hopman & MacDonald, 2018). However, the gold standard for grammar learning, as for any type of regularity learning, is generalization: the ability to apply the regularity to novel instances. In this dissertation, I want to test whether production training also improves generalization to novel instances.

First, I review the literature on generalization in language learning from the point of view that production has meaningfully different task demands than comprehension, and that these differences are relevant for theorizing about generalization. In order to do so, I draw on three fairly distinct areas of published literature (Figure 1). Note that while 'training' and 'testing' are explicitly separated in Figure 1, I do not believe that they are inherently different to a learner. Pragmatically, the literatures covered in the two boxes cover different research questions. In the production versus comprehension training literature, focus is on the effects of the two different tasks, so the two tasks are contrasted explicitly in experiments. Very different research questions are central in the extensive literatures on phenomena like the English past tense and language

change – I will mainly sample papers that happen to have tested their questions in both production and comprehension tests.

Figure 1

Structure of the Literature Review

Production vs. Comprehension	Language learning literature	Parallel effects in non-language literature
Training	• Second language acquisition • Infant and child productions • Sound-level learning	• The testing effect • The production and generation effects • Drawing as object production • …
Testing	• English past tense & plural • Retreating from overgeneralization • Language change	

Defining Generalization

Before reviewing the literature, it is important to establish what I mean, and what I don't mean, with the word 'generalization'. I define generalization here as *the ability to apply a regularity to novel instances.* I believe this sense of generalization is related to the transfer-sense of generalization as it is often used in e.g. the memory literature. For example, Pan and Rickard explicitly define transfer as "the productive use of prior learning in a novel context […] any situation that is different in some way from that in which original learning took place [….] a different topic, a different goal, a different test type, or any number of contextual changes" (2018, pp. 710-711). One might argue that transfer of one situation or task to another isn't the same as generalization of a grammatical rule to a novel lexical item. I think they are related – in

both situations, there's a learned regularity, and it is applied in a novel setting. Woollams et al. use the word 'generalization' for both senses: they use the applying-a-regularity sense "generalised grammar" (2009, p. 56) as well as the transfer sense "our study indicates that conclusions concerning the mechanisms involved in inflectional morphology drawn from performance in standard form-based elicitation tasks do not necessarily generalise to the process underlying past-tense generation from meaning" (p. 73). I will differentiate the two senses in this dissertation proposal by using 'generalize' only for the applying-a-regularity sense, but I do believe this is related to 'transfer' (generalization) to different situations, tasks and mechanisms.

Within the applying-a-regularity sense of generalization, it depends on the level of language processing that is of interest what a 'novel instance' is. For example, at the phonological level a child may say 'nana' instead of 'banana', thus overapplying the regularity that most words she knows start on a stressed syllable. At the lexical-semantic level, a child might say 'dog' while pointing to a sheep, thus overapplying a category to include more of the semantic space than it does. At the morphological level, a learner may come up with forms like 'runned', and at the syntactic level, with forms like 'explain me this', again overextending known regularities. In order to emphasize that I think about generalization (and generalization errors) at these different levels as stemming from the same underlying process, I will use the word generalization interchangeably for all of these levels (in line with e.g. Ambridge et al., 2013; MacDonald, 2013a).

The flip side of generalization is overgeneralization, and again authors differ in how they use or don't use this word. The classic example of an overgeneralization is when learners produce forms like 'runned', where they extend a regularity beyond its attested use. However, due to the relative sparsity of spontaneous errors like these, overgeneralizations are often studied

with grammaticality judgment tasks. Some authors use ‘overgeneralization’ only for production tasks, and carefully describe participants’ endorsements of overgeneralizations in comprehension tasks like grammaticality judgments as “the rated acceptability of overgeneralization errors” (Ambridge, et al., 2012, p. 263). Other authors do use ‘overgeneralization errors’ to describe participants’ behavior in both comprehension and production tasks – for example, in judgment studies “children are relatively more willing to overgeneralize infrequent than frequent verbs” (Goldberg, 2016, p. 372) and “the results from production and comprehension are complicated by a bias […], so that over-generalization is more apparent with ….” (Wonnacott et al., 2008, p. 188). Yet other articles talking about similar phenomena avoid using either the word generalization and overgeneralization, and simply refer to ‘attested uses’ (e.g. Robenalt & Goldberg, 2015).

In this thesis proposal, in line with several authors I cite (e.g. Goldberg, 2016; Wonnacott et al., 2008), I will use ‘overgeneralization’ to describe *any extension of a regularity beyond its attested use, independent of the task modality.* Thus, in addition to a participant producing a form like ‘runned’, I will refer to judging a form like ‘runned’ to be grammatically correct as an overgeneralization error, and to incorrectly attributing a member of a less common neighborhood (e.g. neuter words in Dutch, taking the ‘het’ determiner) as belonging to a more common neighborhood (e.g. masculine words in Dutch, taking the ‘de’ determiner) as an overgeneralization error. I choose to use the word overgeneralization in this task-general way because, while I do believe task differences matter for when learners are likely to make errors like these, in terms of learning a regularity I consider them as errors of an equivalent type.

I view generalization, applying regularities to new instances, and overgeneralization, where this application extends beyond attested uses, as flip sides of the same coin. To a learner,

both a succesful generalization and an infelicitous overgeneralization simply are attempts of applying a regularity to a novel instance. The distinction might be immediately clear to a fluent user, and might be clear to the learner afterwards, upon receiving feedback, but I don't believe they are different in the moment of making the attempt. This is in line with other authors who describe unattested uses as both generalization and overgeneralizations (e.g. Wonnacott et al., 2008, p. 183). I will refer back to this later, when identifying patterns of (over)generalization as dependent on modality in Section Testing Production Versus Comprehension in Language Learning.

Finally, I'll review some literature in which the word 'regularize' is used. In these studies, learners over-apply a regularity (compared to its use in the input), but instead of scoring this as overgeneralization errors, the researchers score it positively as a sign that the learner is making the language more systematic (e.g. Hudson Kam & Newport, 2005; 2009; Hendricks, Miller & Jackson, 2018). Thus, overgeneralization and regularizations are both cases in which users apply a regularity beyond its attested uses, and typically the only difference is in how the experimenter values these uses. Confusingly though, since the term 'regularize' became common in the literature, 'overregularization' is now used as a synonym for 'overgeneralization' (e.g. Ramscar et al., 2013, p. 760; MacDonald, 2013a, p. 7). In the experiment proposed here, I will use neither 'regularize' nor 'overregularize', since I am interpreting any overextension as an error rather than a way to make the language more systematic.

Language Production Versus Comprehension Training

Why am I interested in contrasting the learning consequences of production and comprehension? Language production is different from language comprehension in several ways that may affect learning. Production is harder than comprehension and requires more attentional

resources (Boiteau et al., 2014), which might in turn increase depth of processing and thus learning during production. Language production and comprehension typically draw on different memory processes: production often involves recall whereas comprehension involves recognition. Recall practice has in turn been shown to lead to better learning than recognition practice (Roediger & Karpicke, 2006). People also tend to remember words they themselves say better than words they hear another person say (MacLeod & Bodner, 2017; Hoedemaker et al., 2017; Yoon et al., 2016). Finally, language production involves making task-relevant choices, which for other motor tasks has been shown to improve learning (Carter & Ste-Marie, 2017). I will revisit these differences and more in the third section of this review to see whether there is evidence that they may impact not just learning but also generalization.

Based on these differences between production and comprehension, we hypothesized that language production training might be a stronger learning experience than language comprehension training (Hopman & MacDonald, 2018) – not just for single words, as had been shown in memory studies (Karpicke & Roediger, 2008) but also for grammatical dependencies between words. To test this, we designed an artificial language learning experiment that contrasted production and comprehension training in a between-subjects design that carefully balanced attention and other less interesting task demands. We found that, even when controlling for individual vocabulary knowledge, participants with production training outperformed participants with comprehension training on grammar comprehension tests.

While this is a promising result, participants in this study were not truly tested on generalization. The grammatical dependency of interest was a gender-like suffix agreement on nouns, adjectives and verbs that was deterministically dependent on the type of monster the sentence described. Critically, participants were tested on novel combinations of familiar

elements (e.g. a different sentence about a familiar monster), but never on novel elements that would have required generalizing the suffix use to new vocabulary words. Given that production participants did better on novel combinations of familiar elements, I expect that production participants would also outperform comprehension participants on true generalization trials, but we have not tested this yet.

These results, while expected based on single word memory findings, were surprising in the light of the second language acquisition literature. A meta-analysis in that literature found that, especially when measured within one week, comprehension performance benefits more from comprehension than production instruction (Shintani et al., 2013). A critical difference, that we believe explains why we found such different results, is the way in which production training was implemented in our experiment. Second language acquisition experiments have typically not tried to equate task demands, and have often implemented production as simply reading out loud or repeating a phrase given by a teacher (e.g. Macdonald et al., 1994). Again drawing on the memory literature, there is evidence that for learning single words, retrieval practice that involves generating the word from memory is more effective than simply repeating a teacher (Kang et al., 2013). Tasks like repeating a teacher or reading out loud don't involve generating the to-be-learned language from long-term memory in order to plan and produce a sentence, and are thus missing precisely the elements that we hypothesize make production a stronger learning experience. Thus, while the second language acquisition literature has often tested comprehension versus production and theorized about it, these results are less relevant because of the unbalanced way in which the contrast between production and comprehension was implemented.

There are some hints that, in naturalistic second language learning, production might be a stronger learning experience than comprehension. De Wilde et al. (2019) investigated which type of experience with English is most predictive of Flemish kids' level of English at the start of formally learning the language in school. They found evidence that interactive types of experience (e.g. gaming, using social media in English) were more predictive of level of English proficiency than passive types of experience (e.g. reading and watching television in English). Of course, one important difference with passive exposure is that these interactive types of exposure include production. Furthermore, language learners who are immersed in a foreign language context learn that language much better than non-immersed learners (Barik & Swain, 1978; Fortune, 2012; Hartshorne et al., 2018) – and one big qualitative difference between immersion and non-immersion contexts is that learners are forced to speak the foreign language frequently. This benefit of immersion learning, among other findings, led to the proposal of the 'output hypothesis' in second language acquisition which, contrary to mainstream theories at the time, focuses on ways in which production can be helpful for learning a second language (Swain, 2005; for criticisms, see Krashen, 2003).

It is well-established that in infants, comprehension, referred to as perception at the sound level, shapes production. Infants' babbling shows influences of the language they hear in their environment (e.g. Goldstein & Schwade, 2008), and children learn to speak the language they hear around them, even when they are adopted into a country with a language very different from the language they are initially exposed to (Pallier et al., 2003). Generally, research on children's language development has overwhelmingly focused on the importance of the input children are exposed to, like the '30-million-word-gap' (Hart & Risley, 1995). However, some more recent evidence points to children's own opportunities to produce language as potentially an even more

important predictor of language learning outcomes. For example, the amount of conversational turns a child has with their caregivers is a stronger predictor of language outcomes than the amount of input the child receives (Zimmerman et al., 2009). In even younger babies, there is evidence for a social feedback loop, where speech-like sounds the infant makes trigger parental response (Warlaumont et al., 2014). There is experimental evidence that babies can learn from such feedback that is time-locked to their own productions (Goldstein & Schwade, 2008), pointing to an important role for early production (babbling). Furthermore, spontaneous imitation of caregiver speech by infants predicts vocabulary at a later age (Masur, 1995).

Finally, children with SLI tend to have concurrent motor difficulties (Sanjeevan et al., 2015); if motor skills are impaired in these children, this would impact their language production skill more than their language comprehension skill. Speculatively, if language production indeed plays an important and different role in language learning from language comprehension, then motor impairments affecting production skill might have relatively big downstream consequences on general language learning ability. Thus, there is plenty of suggestive evidence throughout early language development that language production might play a more important developmental role than it is usually thought to do, pointing to the need to move beyond just measuring and trying to improve only the *input* that a child receives.

Direct evidence for the claim that production can also shape perception in young infants is more recent. A teether toy that impairs tongue movement can influence perception of sound contrasts in 6 month old infants, showing a sensorimotor influence on speech perception (Bruderer et al., 2015; Choi et al., 2019). Furthermore, there is evidence that the sounds an infant can produce stand out in the input and are processed differently, because of the richer motor processing associated with those sounds (articulatory filter hypothesis; Vihman, 2017; DePaolis

et al., 2011). While these studies don't speak to generalization immediately, they are in line with accounts in which production experience provides a different and strong learning experience with downstream consequences for comprehension (MacDonald, 2013a).

With regards to generalization, Vihman (2017) has noted that children sometimes seem to form a preferred template for production based on some early produced words, which they then overapply to words they subsequently learn to produce. She provides the example of German learning infant Annalena, who starts off with many words consisting of a duplicated syllable, and overgeneralizes that pattern to e.g. 'baba' for 'bauch' (belly). It would be interesting to know whether children would accept these overgeneralizations in comprehension if e.g. their parent, who normally says 'bauch', says 'baba' to refer to belly. At the sound level, children can hear distinctions that they themselves cannot produce yet (e.g. 'fis' and 'fish', Berko & Brown, 1960), and do not accept similar mispronunciations from adults. Thus, it's possible that these overgeneralized templates are specific to production, and might be caused by production difficulties (MacDonald, 2013a) – specifically, overgeneralized forms like 'fis' and 'baba' may simply be easier for the child to produce.

At the sound level, production practice with adults has had mixed results. Baese-Berk and Samuel (2016) found that production training impairs learning of a non-native sound contrast, whereas perception training improved this learning. Bixby (2017) studied several different sound contrasts and found that production practice improved perception only in the cases where participants had started to be able to distinctly produce the different foreign speech sounds. This might explain the difference in results with our morphology-focused study (Hopman & MacDonald, 2018): in our study, the phonetics of the artificial language was purposefully simple, and virtually all production participants managed to produce the different words and their

grammatical suffixes successfully on at least some trials. It has also been hypothesized elsewhere that perception and the perceiver's needs are most important at the level of prosody and pronunciation, whereas production and the producer's needs might play a more important role at the lexico-syntactic level (MacDonald, 2013b). Regardless, while the effect of production versus perception training on generalization has been investigated at the sound level (e.g. Bixby, 2017) the mixed results as well as the possibility that the perceiver's and producer's needs play different roles at the sound level make it hard to draw conclusions or make predictions for the grammar level.

In summary, much of the existing research that directly contrasts language production and comprehension training either doesn't speak to generalization (e.g. Hopman & MacDonald, 2018) or isn't informative because of mixed results (e.g. Bixby, 2017) or unbalanced experiments (Shintani, Li & Ellis, 2013). That being said, there are theories drawing on experimental evidence in both infants (Vihman, 2017) and children and adults (MacDonald, 2013a) proposing that language production plays a special role in language learning with downstream consequences for language comprehension. In line with this, one might expect language production to also improve generalization for regular forms and to potentially lead to more overgeneralization than language comprehension.

Testing Production Versus Comprehension in Language Learning

I now turn to three related areas of the grammar learning literature: studies on the English past tense and plural, studies on how humans retreat from overgeneralization and studies on language change. Each of these three areas has their own research questions, theoretical debates, and variables of interest they typically manipulate in experiments, and could merit a review on their own. Here, rather than try to review these areas in full, I mostly sample papers that have

tested both comprehension and production, as well as some papers that are typical for how each area has been approached.

The English Past Tense and Plural

A well-studied example of overgeneralization at the morphology level is the English past tense. Regular verbs in English get the suffix '-ed' to indicate past tense (e.g. *'walk*', '*walked*'). Many frequent verbs in English are irregular and have an idiosyncratic past tense (e.g. '*eat*', '*ate*'). However, learners of English, both children and second language learners, will occasionally overgeneralize the regular past tense rule and apply it to irregular verbs, creating overgeneralized forms like '*eated*' by simply adding '*-ed*' to the present tense form of the irregular verb.

Overgeneralizations are typically thought of as a language-production phenomenon. Overgeneralizations are present in learners' speech, and since adult native speakers don't tend to produce them, it doesn't seem meaningful to think about them in comprehension at all. That being said, overgeneralizations are rare in spontaneous speech, making it hard to study them. In order to investigate generalization in language learning, Berko (1958) developed the famous wug test. In this elicited production task, children are exposed to a form like the present tense of a novel pseudo-English verb (e.g. '*spow'*), and then asked to finish a sentence that prompts them to generate the past tense (e.g. '*Yesterday he .....*'). Of course, this elicitation task also works to test the learner on the past tense of known verbs. Even in elicited production studies like these, errors are relatively rare, meaning that researchers need to collect a substantial amount of data in order to have enough errors to analyze.

In addition to these elicited production tasks, which are effortful for the child and the researcher both, grammaticality judgment tasks are often used to test children's judgment of

overgeneralization errors. In an example task, a child might hear a puppet say '*Yesterday he eated'* and be asked to judge on a sliding smiley scale, a child-friendly version of a Likert scale, whether the sentence the puppet says sounds good or bad (e.g. Ambridge et al., 2008). An obvious advantage of this is that the researcher can get a learner to respond to any form they are interested in, rather than waiting for the learner to produce that specific form spontaneously. The, often implicit, assumption is that forms that the learner endorses in a grammaticality judgment task are forms that they would produce themselves, and many review papers aggregate data from these different tasks without distinguishing them (e.g. Ambridge et al., 2013; Goldberg, 2016). The data I'll review next will challenge that assumption.

Table 1

Past Tense Production (Kuczaj, 1977) and Comprehension Data (Kuczaj, 1978)

Age (y.)	Grammatical (**ate**)				Base+ed ungrammatical (**eated**)				Past+ed ungrammatical (**ated**)			
	Sp. Pr.	Elic. Pr.	Forc. Ch.	Acc. Jud.	Sp. Pr.	Elic. Pr.	Forc. Ch.	Acc. Jud.	Sp. Pr.	Elic. Pr.	Forc. Ch.	Acc. Jud.
3-4	77%	71%	71%	99%	15%	29%	28%	77%	9%	0%	1%	25%
5-6	-	51%	19%	98%	-	42%	18%	52%	-	7%	63%	80%
7-8	-	99%	98%	100%	-	1%	2%	19%	-	0%	0%	4%

Note. Sp. Pr.: Spontaneous Production; Elic. Pr.: Elicited Production in a version of the wug-task. Forc. Ch.: Forced Choice. Acc. Jud.: Acceptability Judgment; both this and Forced Choice are considered comprehension tasks; y..: years old.

I'll review Kuczaj's (1977, 1978) past tense data here (see summary in Table 1). There are several reasons I focus on these data. First of all, the two main competing theories in the past tense debate both cite and draw on these data (e.g. Marcus et al., 1992; Rumelhart & McClelland, 1986). Second and more importantly, they provide a thorough comparison of English past tense performance by different age groups on several different comprehension and

production tasks. I will identify three patterns of results in these data that we will see repeated throughout the rest of this section. Note that, as explained earlier, I view succesful generalization and overgeneralization errors as two sides of the same coin, – thus, I expect to see similar patterns for succesful generalization as for erroneous overgeneralization (see section Defining Generalization).

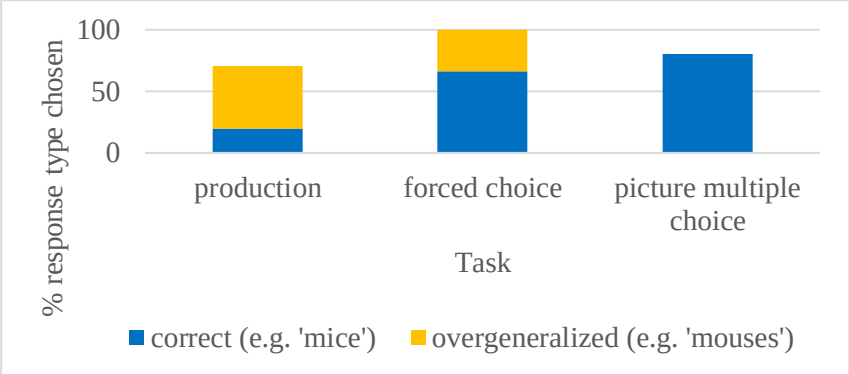
Pattern (a): Overgeneralization Errors Are a Production Phenomenon. On the surface, the data in Table 1 (focus on numbers highlighted in green) seem to match the classic narrative of English past tense acquisition. Production in an elicitation task shows a U-shaped pattern across ages, with young, 3-4 year old kids initially producing mostly correct irregular past tense forms like 'ate', followed by a period in which 5-6 year olds' performance on irregulars drops off, until they are at ceiling around 7-8 years old (leftmost green column). Meanwhile, overgeneralized erroneous forms like 'eated' go through an inverse U-shape, showing that that's what kids are producing instead of the correct form (rightmost green column). Finally, kids in all of these age groups judge correct irregular past tense forms like 'ate' as correct (>98%), irrespective of where their production is at, confirming that overgeneralization errors are a production phenomenon (yellow column).

A related example is data for the English plural. Ramscar & Yarlett (2007) find in a behavioral experiment that 3-5 year old children's comprehension of the plural is above chance and better than their production (Figure 2). Children score on average 66% correct in a forced choice comprehension task between the correct irregular plural and an overgeneralized plural, and 80% correct in a multiple choice comprehension task where they choose between pictures of several singular and plural objects. In contrast, they only produce correct irregular plurals on 20% of the trials, compared with 51% erroneous overgeneralized plurals (e.g. *'mouses'*). We'll

see this pattern, of more overgeneralization errors in production than comprehension tasks, come back across different areas of the literature.

Figure 2

Example English Plural Data



Note. 3-5 Year old children's performance on the English plural in an elicited production task and two different comprehension tasks as reported in Ramscar & Yarlett (2007). Note that due to the different nature of the three tasks, each task had different types of possible responses. Only the two most common response types across tasks are relevant for this review and depicted here.

Pattern (b): Task Demands Matter Beyond Production Versus Comprehension.

However, looking at the rest of Kuczaj's data in Table 1 more closely reveals that the narrative of overgeneralization as a pure production phenomenon is too simplistic. At 5-6 years old, the bottom of the classic U-shape, the results of the forced choice comprehension test (grey highlighting) show a striking pattern: given a choice between three forms like 'ate', 'eated' and 'ated', kids at this age prefer erroneous hybrid forms like 'ated', despite barely ever producing these forms themselves and despite rating correct irregular forms higher on acceptability judgments. This is our first example that sometimes, two different comprehension tasks show different patterns of accepting overgeneralization errors. Similarly, there is evidence that adults

make overgeneralization errors according to a frequency by regularity interaction in a task where they generate the past tense from the present tense stem, but not in a task where they generate the past tense from a picture (Woollams et al., 2009). Thus, different production tasks can also show different patterns of results, showing again that task demands matter in a way that is more subtle than just production version comprehension.

Pattern (c): Evidence of Overgeneralization in Production *and* Comprehension Tasks. Finally, it is fairly common in the literature to only measure generalization performance in either a production or a comprehension test. For example, Ramscar et al. (2013) see overgeneralized plurals in a picture-elicited production pretest for both 4- and 6-year old children. I note here as well that, while Ramscar and Yarlett conclude from their data that "children who over-regularize plurals in production nevertheless have representations of the correct adult forms in memory" (2007, p. 940), I myself would draw different conclusions from their data. While the reported scores of 66% and 80% correct in the two comprehension tasks are significantly above chance and significantly better than the number of correctly produced irregular plurals, my guess is that the 33% choices of erroneous overgeneralized forms like 'mouses' is significantly above 0% (Figure 2). Thus, I would conclude that we can see evidence of overgeneralization errors in *both* comprehension and production tests, albeit significantly more in the latter.

Now that I have identified three patterns of results when both production and comprehension are tested based on the English past tense literature, I will review literature on retreating from overgeneralization and language change, to see whether studies in those fields have shown similar patterns of results.

Retreating From Overgeneralization Errors

While children make overgeneralization errors like '*eated*', they eventually learn to express meanings like these in the same way that adult native speakers do. The learning strategies people can draw on to retreat from overgeneralization errors while at the same time maintaining the ability to generalize rules when appropriate (a problem known as Baker's paradox) has generated a vast amount of research (see Ambridge et al., 2013 for a review). Here, I again mainly sample papers that have tested both production and comprehension, and in doing so we will see several of the patterns identified in the past tense and plural literature repeated.

Wonnacott and colleagues did several experiments to determine how adults and kids generalize based on distributions in the input, and tested participants afterwards on production and forced choice comprehension. Furthermore, they tested child participants in an act out comprehension task and adult participants in a grammaticality judgment task. Interestingly, they show evidence that adults are more willing to generalize in their own productions and forced choice judgments than in their grammaticality judgments, which were more conservative, staying closer to the input (Wonnacott et al., 2008). In this, we see patterns (a), more generalization in a production than in a comprehension task, and (b), not all comprehension tasks show the same patterns of overgeneralization, repeated.

Children, like adults, show strikingly similar patterns of generalization in both the production and forced choice comprehension tasks, and actually show a similar pattern of results in their third task, act out comprehension (Wonnacott et al., 2013) – this is an example of pattern (c), similar generalization in production and comprehension tasks. Likewise, Perek and Goldberg (2015, 2017) conducted several studies looking at the role of construction meaning in generalization behavior with adults, and found that analyzing the production tasks and the

comprehension tasks (in this case, grammatical judgments) led to similar conclusions about how adults generalize.

One well-researched example of a learning strategy that can help solve Baker's paradox is pre-emption, in which a novel, overgeneralized form is pre-empted by a different form that means the same but is available in the input. Goldberg (2016) proposes competition-driven learning and prediction as mechanisms for pre-emption in native speakers. If a listener is presented with forms like '*He made me giggle'* whenever they are predicting a form like '*He giggled me.*', there will be a prediction error, which will in time teach the listener to expect (and use) the attested form. Evidence for the use of pre-emption as a learning strategy comes mainly from grammaticality judgment tasks in comprehension-only studies. For example, adult native speakers of English rate novel, unattested uses of English verbs with a competing alternative as less acceptable than novel uses without a competing alternative (Robenalt & Goldberg, 2015). Thus, the overgeneralization '*He slept the afternoon away*.' is more acceptable to native speakers than the overgeneralization '*He giggled me*'.

Non-native speakers do not differentiate in their ratings between novel uses with or without a competing alternative (Robenalt & Goldberg, 2016) – evidence of overgeneralization errors endorsed in studies that test only comprehension, fitting with pattern (c). This is in line with other work showing that even highly fluent non-native speakers do not engage in predictive processing, at least at the grammar level, to the same degree that native speakers do (e.g. Grüter et al., 2012). Within Goldberg's framework, if non-native speakers do not predict the competing alternative, there is no prediction error for them to learn from and unattested (overgeneralized) forms are not unexpected. In a similar study, Tachihara and Goldberg (2019) show production evidence for overgeneralization in non-native speakers and hypothesize that this is because non-

native speakers make less use of pre-emption, because they don't predict during online comprehension to the same degree that native speakers do – a comprehension explanation for a production phenomenon.

Language Change

This same overgeneralization phenomenon is studied under a different name, overregularization, to answer questions about language change. In this literature, the overgeneralization 'errors' are seen as a consequence of learnability biases that make a language more regular. Hudson Kam and Newport's (2005) seminal paper is the first well-known example of a language learning experiment that directly sought to investigate language change. In this study, participants learned an artificial language which contained unpredictable variation in determiner usage in the input. After learning, participants were tested on their determiner use in a sentence completion test. Participants also completed a grammatical judgment test, where they had to rate novel sentences.

This experiment showed that adult participants probability matched with the input, whereas children regularize the input. So, in the production task, the proportion of nouns that adults produced with determiners roughly matched the percentage in their input, whereas most children either chose to always produce determiners, or to never produce them, thus making their own output more systematic than the input language they were exposed to (by making overgeneralization 'errors'). The grammaticality judgment task showed a similar pattern of results to the production task for each of the two age groups – thus, this study falls under pattern (c), with similar levels of overgeneralization in production and comprehension tasks. In a follow-up study, the authors wanted to see whether, given even more variable input, adults would look more like children and regularize the language (Hudson Kam & Newport, 2009). This variable

input made it virtually impossible for adults to fully probability match, and in both the production and the grammaticality judgment task adults show evidence of regularizing the input: the more variable their input, the more adults regularize (by making overgeneralization 'errors'). This again falls under pattern (c), with the production and comprehension tests showing a similar pattern of results.

The authors also do a simpler variant of this second experiment with kids. Interestingly, in this experiment, kids show evidence of overgeneralization errors in their productions but not in a comprehension task, in line with the classic view of overgeneralization errors as a production phenomenon (pattern a). In a much smaller scale study, Schwab et al. (2018) focused on explicitly contrasting comprehension and production tests. Children are exposed to a small language that takes different verb modifiers for masculine versus feminine nouns. When tested on production, they make many overgeneralization errors, but when tested on comprehension they do not accept these errors, similar to the children in Hudson Kam and Newport (2009), and again in line with pattern (a), overgeneralization errors as a production choice.

Conclusions From Research Testing Production Versus Comprehension

While I have reviewed all studies in this section through the lens of results in production versus comprehension tests, it is worth noting that in almost none of these studies (except Schwab et al., 2018), comprehension versus production was explicitly manipulated as the contrast of interest. An important consequence of that is that, since it was not explicitly manipulated, production and comprehension tests were not balanced for task demands. Furthermore, of course all of these studies did manipulate other contrasts, informed by the central questions and theories of those literatures – for example, the different roles children and adults might play in language change (Hudson Kam & Newport, 2005; 2009). Together, these

two reasons may explain why the results of production and comprehension tests in various language learning studies seem fairly disparate, at least at first glance. I identified the following three main patterns in the results:

(a) More (over)generalization errors in production than comprehension tests (e.g. Schwab et al., 2018), in line with the view that overgeneralization by learners could, at least in part, be stemming from task differences between production and comprehension (MacDonald, 2013a).

(b) The importance of looking at task demands for (over)generalization beyond a simple production versus comprehension distinction. We have seen that sometimes, different production tasks (e.g. Woollams et al., 2009) show different patterns of overgeneralization errors, and sometimes, different comprehension tasks (e.g. Kuczaj, 1978) show different patterns of accepting or choosing erroneous overgeneralized forms.

(c) Finally, it is common to find similar patterns of results for (over)generalization in both comprehension and production tests (e.g. Hudson Kam & Newport, 2005), and to find (over)generalizations in one of the two when only one is tested (e.g. Robenalt & Goldberg, 2016 for comprehension; Ramscar et al., 2013 for production).

Perhaps more telling is the pattern of results that is absent here: evidence of more (over)generalization errors in comprehension than production. A comprehensive meta-analyses of each of the three mentioned literatures is beyond the scope of this introduction, but it is suggestive that in specifically sampling studies that tested both production and comprehension, I found no such results. Thus, my conclusion of the behavioral evidence reviewed across these three different areas of language learning research is that while overgeneralization errors happen

in both comprehension and production, if anything they happen more in production than comprehension, and certainly not the other way around.

Comprehension- and Production-Like Training in Non-Language Research

Because of this limited informativeness of the language learning literature, I now turn to other areas of psychology where some contrasts have been studied that resemble production versus comprehension training. These are not language learning studies (at most, they are about learning new words), and so the evidence they provide is merely suggestive for potential differences between production and comprehension training. As we will see, these studies speak mostly to how production might affect generalization – overgeneralization, the flip side of the coin where applying a learned regularity leads to errors, is either not relevant in these literatures or has not been studied extensively yet.

The Testing Effect and Retrieval-Based Learning

In education and memory research, an important and well-studied contrast is between recognition and recall training. Recall training, also called retrieval training, is any practice activity with to-be-learned materials that requires the student to generate the materials from memory. A large body of research has convincingly shown that, compared to a restudying condition, where students get to read the materials several more times, any condition that involves retrieving the materials from memory improves learning (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006; see Adesope et al., 2017 for a meta-analysis). There is a clear parallel here with production versus comprehension training: production generally involves memory retrieval whereas comprehension involves recognition.

The typical control condition in these memory experiments, restudying, resembles passive comprehension. From a cognitive perspective, these two tasks aren't very well-

controlled: the retrieval condition involves an active task with overt responses whereas the control condition doesn't require overt responses. That being said, students often choose to reread materials as their main mode of studying for exams, so the more passive control condition is very ecologically valid. An interesting parallel here is that foreign language students often find it scary to speak in the language they are learning and may avoid speaking for this reason (Young, 1990), despite production potentially improving learning compared to comprehension practice.

However, some education studies have used active comprehension tasks as well, e.g. multiple choice tests. Interestingly, this literature has also noted task-dependent results. For example, production-like short answer tests generally lead to better learning than limited retrieval fill-in-the-blank tests or comprehension-like multiple choice tests (Hinze & Wiley, 2011; Kang et al., 2007). Furthermore, a note-taking control condition (another form of production) provides an equivalent learning boost to a free recall condition (McDaniel et al., 2009).

As to generalization, there is some evidence that testing, compared to rereading, leads to better performance on near transfer questions (Butler, 2010; McDaniel et al., 2009). Other studies that did not find transfer (e.g. Tran et al., 2015) are again accounted for by noting important task differences: when they are redone with a different task format (e.g. simultaneous presentation of facts instead of one by one presentation), transfer is found (Eglington & Kang, 2018). A recent meta-analysis estimated the effect size of transfer for inference questions at d = 0.32, with a 95% Confidence Interval of [0.085, 0.56], establishing that there is moderate transfer on educational materials (Pan & Rickard, 2018, Table 2). While the methods and materials in these studies are fairly different from typical language learning experiments, at least these results

suggest that production, which involves retrieval, might indeed help benefit transfer, which I view as akin to generalization.

Nearly all of the memory research I reviewed here draws on verbal materials and tasks. Even when the material to be learned is e.g. biology, the multiple choice and short answer test formats used in these experiments rely entirely on language processing. Unfortunately, though some views of memory do explicitly cast verbal short term memory as emergent from the language production and comprehension processing system in action (e.g. Schwering & MacDonald, 2020; MacDonald, 2016), most mainstream memory theories do not draw on psycholinguistics. For example, much theorizing about the testing effect focuses on the number and distinctiveness of different routes to retrieval (e.g. 'study memory' and 'test memory', Rickard & Pan, 2018). These accounts do not easily lend themselves to integration with prevailing psycholinguistic theorizing about how language production and comprehension work.

Theories that combine psycholinguistics with memory would be mutually beneficial here. Our own work is an example of psycholinguistics benefitting from the vast amount of experimental results available in the memory literature (Hopman & MacDonald, 2018). Conversely, given that the majority of these memory experiments draw on verbal material, it stands to reason that bringing what we know about language processing into theorizing about these memory processes would further our understanding of those processes (see e.g. Zormpa et al., 2019).

The Production and Generation Effects

The production effect and the generation effect, two other effects from the memory literature that might contribute to a better memory for produced than comprehended items, have limited applicability to generalization differences between the two. The production effect

(MacLeod & Bodner, 2017) holds that within a list, people remember words that they spoke out loud better than words they only heard or read. Theoretical explanations for this production effect hinge on *comparative* distinctiveness of item-specific memory traces. Similarly, the generation effect (Begg & Snider, 1987) is another within list memory effect, with people better remembering words they generate (with the first two letters given) compared with words in this same list that were read normally. This turns out to be an artifact of the unusualness of the generation task compared to normal reading, again temporarily increasing the comparative distinctiveness of the item's memory trace rather than changing anything about the item's representation in long term memory. Both of these effects are more easily found in mixed lists and fade away in blocked (between-list) designs (Bertsch et al., 2007; MacLeod & Bodner, 2017). Because of the focus of these experiments on within list effects and of these theories on items' comparative distinctiveness, these literatures cannot currently speak to generalization. Both of these effects may reflect isolated, small components of why naturalistic language production and free recall testing have a big effect on learning.

Drawing as Production of Object Representations

Practice drawing (visually producing) objects improves people's recognition of those objects (Fan et al., 2018), an interesting parallel with how producing language improves comprehension (Hopman & MacDonald, 2018). The theoretical emphasis in this work on how drawing objects might change object recognition is on how memory representations can change by accessing them, and how accessing them in the service of different tasks may change the memory representations in different ways (Fan et al., 2018). While the drawing work and its emphasis on the role of tasks is relatively new, the idea that memory representations can change by accessing them, especially through language, is well-established (e.g. Loftus & Palmer, 1974;

for a recent review see Zaragoza et al., 2006). Since the drawing work is explicitly inspired on analogy to language production and comprehension, the (developing) theories around it lend themselves naturally to application in psycholinguistics.

Explanation

Explaining out loud why an exemplar is part of a category that is new to a learner fosters the learner's ability to generalize (as well as the propensity to overgeneralize) category membership (Williams & Lombrozo, 2010; Williams et al., 2013). Explaining out loud helps people generalize because it focuses attention on the relationships between the different items; similarly, we have hypothesized that speaking a new language out loud helps people learn the regularities because it increases binding between the different elements in a phrase or sentence (see Figure 1 in Hopman & MacDonald, 2018). These explanation experiments have found benefits of explanation (production) in contrast with both passive comprehension-like silent study conditions and other production conditions (e.g. describe exemplars), again highlighting that not all production tasks are alike (pattern b).

Attention, Depth of Processing, Active Learning, Motivation and Anxiety

Finally, a relevant difference between comprehension and production that might contribute to learning and outcome differences between the two modalities is the amount of attention typically required. Production tasks typically require more attention (Boiteau et al., 2014) and involve greater depth of processing than comprehension tasks. Greater depth of processing is associated with better learning and generalization of materials in second language learning contexts (Oded & Walters, 2001). In a similar vein, production under typical circumstances is a more active task than comprehension – for example, it involves making task-relevant choices like which words to use. Task-relevant choices have been shown to increase

learning and generalization in non-language motor tasks (Carter & Ste-Marie, 2017; Sanli et al., 2013). One dominant account of the benefit of making task-relevant choices in motor learning tasks, called the information processing account, posits that it is the estimation of errors while learners wait for feedback that drives the learning benefit (Carter & Ste-Marie, 2017). This is reminiscent of error-driven learning in computational models of language learning (e.g. Plunkett & Marchman, 1991 for an example modeling the past tense of English). More generally, active learning strategies and opportunities are associated with better learning of, for example, categories (Markant & Gureckis, 2014; Sim et al., 2015; see Zettersten, 2018 for a review).

However, it is important to note for all of the research mentioned in the previous paragraph that it is possible to disentangle the production versus comprehension contrast of interest in this review from both active versus passive learning and deep versus shallow processing. It is possible and even common in second language learning research to create active comprehension tasks that target deep processing of the foreign language (VanPatten, 2004). Similarly, it is possible to reduce processing demands in production tasks – e.g. a task in which a student reads vocabulary or a passage in a foreign language out loud does not involve active learning or deep processing, especially at the lexical-semantic and grammar levels (see Stroh, 2012 and works cited therein for examples of such tasks in classroom situations). So, task demands and the specific production or comprehension tasks used in classrooms and experiments matter hugely. That being said, in more naturalistic language learning situations in which a learner has a conversation, production and production planning are more attention-demanding and thus potentially more beneficial for learning and generalization than comprehension.

While so far I have discussed production as being inherently more attention-demanding and harder than comprehension, there are other potential differences between production and

comprehension related to this: differences in anxiety and motivation. Speaking in a foreign language is harder than listening to it, and many classroom students experience speaking anxiety (Ansari, 2015; Young, 1990). This anxiety is thought to hamper learning, and this is a reason for some SLA theories to advise against production as a good way for learning a second language (e.g. Krashen, 2003). Interestingly, some research suggests that there is an inverse U-shaped relationship between anxiety and learning outcomes: small amounts of anxiety lead to increased effort and thus performance, and large amounts of anxiety lead to decreased performance (Eysenck, 1979; MacIntyre, 1999). Thus, as long as the anxiety induced in a speaking task is relatively small, increased anxiety may motivate learners to pay more attention or generally put in more effort, leading to an increase in performance (see e.g. Fung & Min, 2016, who showed that a board game involving a speaking task leads to better learning outcomes). While it has been generally established that an increase in motivation leads to better second language learning outcomes (Masgoret & Gardner, 2003), it is less clear how exactly language production, attention, motivation and learning outcomes interact with each other and with generalization.

Conclusions From Non-Language Research

While I see clear parallels between findings that explanation leads to more generalization, and that drawing leads to differences in object recognition with our own finding that production leads to better comprehension (Hopman & MacDonald, 2018), the experimental tasks used are too far removed from language learning to draw strong conclusions from. The observation that language production involves memory retrieval leads to more fruitful parallels. Since there is some evidence that retrieval practice improves transfer, the same might hold for production practice improving generalization in a language learning context.

Summary

I drew on three different areas of literature in this review (Figure 1). First, I reviewed the existing language learning literature that explicitly contrasts production versus comprehension training to see what it tells us about how the two different training tasks might impact (over)generalization (green square in Figure 1). This area of the literature is at once most central to my goal and in some ways least informative due to a relative lack of existing studies and their applicability. Then, I reviewed three related areas of grammar learning research where several published experiments have happened to test generalization and overgeneralization errors in both comprehension and production tasks (blue square in Figure 1). These areas of the literature are extensive, but because production and comprehension were usually not explicitly manipulated, results with respect to this contrast are fairly messy. I found three patterns of overgeneralization results: more overgeneralization in production than comprehension tests, different patterns of results between different comprehension tests or different production tests, and similar results in both comprehension and production tests. From this, I was able to conclude that there is no reason to believe that comprehension tests would lead to more overgeneralization errors than production. Finally, I reviewed other areas of cognitive science research that have studied contrasts similar to the production versus comprehension contrast of interest here (yellow square in Figure 1). Though some of these parallels are speculative at best, from this work I drew evidence that production training may lead to better generalization than comprehension training.

Generalizing Grammatical Dependencies to Novel Lexical Items

Does production training improve generalization of learned regularities to novel lexical items compared to comprehension training? In Hopman & MacDonald (2018), we showed evidence that producing during language learning improves comprehension performance on

comprehension grammar tests. While this is a promising result, participants in this study were not truly tested on generalization. The grammatical dependency of interest was a gender-like suffix agreement on nouns, adjectives and verbs that was deterministically dependent on the type of monster the sentence described. Critically, participants were tested on novel combinations of familiar elements (e.g. a different sentence about a familiar monster), but never on novel elements that would have required generalizing these suffixes to new vocabulary words. In many language learning experiments, testing a learned regularity on novel words is used as gold standard evidence for learning a regularity (e.g. Perek & Goldberg, 2015; Wonnacott et al., 2009; Wonnacott et al., 2008). Thus, I designed this experiment to test whether production practice and comprehension practice differentially affect how learners generalize trained grammatical dependencies when presented with novel lexical items. Following from this, the two main factors of interest that I manipulate are learning condition (production versus comprehension) and item familiarity (familiar versus unfamiliar, the former being trained lexical items and the latter being novel).

Studies of the past tense and similar phenomena in natural languages often show a frequency by regularity interaction (see Woollams et al., 2009 for an in-depth discussion), with learners more likely to make mistakes on infrequent, irregular items. Overregularization in language change studies often takes a slightly different form, with regularities from e.g. a more common gender being applied to nouns from a less common (but still regular) gender (see e.g. Hendricks et al., 2018, for a natural language example of this). In keeping with these latter experiments, I will teach participants to pluralize nouns in different regular categories, with one category being more frequent than the other, even though individual items are all equally

frequent. This is the third factor of interest that I manipulate: neighborhood size (large versus small).

So why might production training help, or hurt, participants when generalizing a regularty to novel lexical items? In finding that production practice improves comprehension performance, we argued that language production has a strong effect on learning especially during the processing involved in planning a sentence (Hopman & MacDonald, 2018). While the learner is planning what to say, elements of the utterance are held together in working memory, which allows binding of the dependencies between the different words to happen. What is unclear from the Hopman & MacDonald experiment is how lexically specific this binding of different types of features is, and whether or not it would extend to production training improving the binding between a category of grammatical features and a set of visual features, *in absence of the specific (lexical) item*. My hypothesis here is that this stronger binding of features relevant to grammatical regularities during utterance planning also increases the ability to correctly generalize. For example, in the artificial language in Hopman & MacDonald (2018), production training could lead to stronger associations between a given morpheme (e.g. '-ok') and certain visual features of the aliens (e.g. multiple legs and eyes). This should then make it easier to combine those same visual features and the same morpheme when processing a novel instance, e.g. a novel scary-looking monster with multiple legs and eyes.

However, in situations where one regularity is more frequent than another, production training could conceivably lead to overgeneralization. For example, in Dutch most nouns get the suffix '-en' to indicate plural (e.g. 'hond', 'honden'; dog, dogs) but some get '-s' (e.g. 'kater', 'katers'; tomcat, tomcats; see also Keuleers et al., 2007). Presented with a novel instance, a production-trained learner may default to the suffix most frequently produced and thus most

closely linked with the plural, leading to mistakes like *‘kateren’ instead of ‘katers’ (tomcats; see Ramscar et al., 2013; Ramscar & Yarlett, 2007). While it may be the case that comprehension-trained learners would make similar overgeneralization errors, it is conceivable that, since production training leads to stronger binding (which may include a stronger association between e.g. the plural and the ‘-en’ ending in Dutch), this tendency to overgeneralize to a default option may also be stronger for production-trained participants.

In order to be able to connect my results to other language learning experiments that have tested both production and comprehension after learning, this experiment will include a production test in addition to several comprehension tests targeting vocabulary and grammar understanding. Since my review of other language learning research indicated that learners are more likely to make overgeneralization errors in production tests than comprehension tests (e.g. Ramscar & Yarlett, 2007), I expect that participants in both learning conditions will commit more overgeneralization errors on the production test than on the comprehension tests.

Note that, when I started designing this experiment, it was unclear how well participants would learn the particular artificial language with the particular training procedures. Another reason for employing several different tests was to have sensitivity to detect condition effects, if any were present. Different tests tended to have different levels of difficulty, I wanted to cast a wide net in case some tests would show floor or ceiling results. Also note that it took extensive piloting, with relatively big changes to the number of aliens, the training procedure and test formats, to get participants to generalize a non-trivial grammatical regularity better than chance level. Many similar studies employ multi-day testing sessions (e.g. Mirković & Gaskell, 2016).

To summarize, in this experiment I will test how well learners generalize learned regularities to novel lexical items, by analyzing patterns of overgeneralization errors in both

production and comprehension tests. This study will contrast production and comprehension in both training and testing. By contrasting these two modalities between subjects in training, I will test whether production practice leads to better generalization on novel lexical items. By conducting both comprehension and production tests after learning, I will build a connection between language learning studies explicitly contrasting production and comprehension training with language learning studies finding different patterns of results when analyzing overgeneralization errors in production and in comprehension tests without explicitly contrasting them. Based on the literature review presented earlier, I hypothesize that production practice leads to more accurate generalization of newly learned grammatical regularities when compared with comprehension practice. More detailed predictions, as well as a possible alternative scenario, are available in Appendix E.

Method

Participants

Participants for this study were recruited through the UW-Madison Psychology SONA extra credit pool, and were prescreened for English as a native language, normal or corrected to normal vision and hearing, and no color-blindness. A total of 641 participants completed the online consent process in qualtrics. Of these participants, 228 completed the experiment and provided fully usable data, and another 36 provided partially usable data. The remaining 377 either did not start or finish the experiment, or their data was unusable for other reasons (see Appendix B).

Materials

Materials for this study consisted of pictures of aliens and an artificial language in both written and recorded audio form that described these aliens.

Artificial Language. Aliens were described with simple noun phrases consisting of a determiner and a noun with a suffix. There were two different neighborhoods, corresponding to the two different visual categories of aliens. Each neighborhood had two unique determiners, one for singular and one for plural, and two suffixes, one for singular and one for plural (see Figure 3; see also Appendix A). In order to create rule frequency differences, there was one large neighborhood (12 familiar and 6 unfamiliar aliens) and one small neighborhood (6 familiar and 6 unfamiliar aliens). Rule frequency differences like these are well-attested in natural languages (e.g. the Dutch plural, see Keuleers et al., 2007), and relevant in this experiment because they allow for overgeneralization errors, where the more frequently encountered rule for the large neighborhood might be incorrectly applied to either the small neighborhood familiar aliens or, especially, the less well learned unfamiliar aliens from the small neighborhood.

Figure 3

Experimental Stimuli

Neighbor-hood	Language		familiar	unfamiliar	Example visual stimuli
	Singular	Plural	18	12	
Large	**dap** roo**zok** **dap** kred**ok** **dap** chag**ok**	**lom** roo**zool** **lom** kred**ool** **lom** chag**ool**	12	6	
Small	**ked** pex**esh** **ked** mip**esh** **ked** bur**esh**	**jeb** pex**aaf** **jeb** mip**aaf** **jeb** bur**aaf**	6	6	

Note. There were two different neighborhoods, each with different suffixes and determiners for both singular and plural. Participants learned about 18 aliens during training ('familiar'), and 12 other aliens ('unfamiliar') were briefly introduced through passive exposure at the end of training, without active practice, in order to test generalization to novel lexical items.

Auditory Stimuli. A list of the 4 artificial language determiners, as well as all 120 possible stem-suffix combinations (nouns) was created and shuffled into three different random orders. A female native English speaker recorded herself speaking all three versions of the 124-word list into a microphone in a sound-attenuated booth. Two native English speakers rated the three recordings of each of the 124 words (grouped by the 4 determiners and 30 stems) for sound quality and consistency of pronounciation. Based on these ratings, I then chose the best overall recordings and normed these for loudness. Noun phrases were then created by pasting together all possible combinations of the 4 determiners and the 120 nouns, with 200 ms silence in between the two words.

Visual Stimuli. There were 2 shape-based categories of aliens (see Figure 3 for examples, and Figure A1 for the full set of stimuli). One category consisted of 18 humanoid aliens, all standing on two legs with two arms by their sides. The other category consisted of 12 arm- and leg-less, wider, blob-shaped aliens. Of these 30 aliens, 12 humanoid and 6 blob-shaped aliens, the 'familiar' group, were actively trained during the training phase of the experiment. In order to test generalization to novel lexical items, six of the aliens in each category were not introduced until the end of training – this subset of 'unfamiliar' aliens was picked randomly for each participant individually. For each alien, there was a singular picture of a single alien, and a plural picture of two identical aliens next to each other.

Training

During training, participants actively learned about the 18 familiar aliens (Figure 3), 12 from the large neighborhood and six from the small neighborhood. Like in Hopman & MacDonald (2018), training alternated between blocks of passive and active trials. There were 12 training blocks. During each block, six different aliens were trained both passively and actively.

Of these six aliens trained per block, four were from the large, and two from the small neighborhood. This worked out so that each familiar alien was encountered during four different training blocks, twice singular and twice plural, for a total of four passive and four active trials per familiar alien.

Each training block consisted of two sub-blocks: a passive exposure sub-block consisting of one passive exposure trial for each of the six aliens trained in that block, and an active training sub-block, consisting of one active training trial for each of the six aliens trained in that block. Trials were randomly ordered within each sub-block.

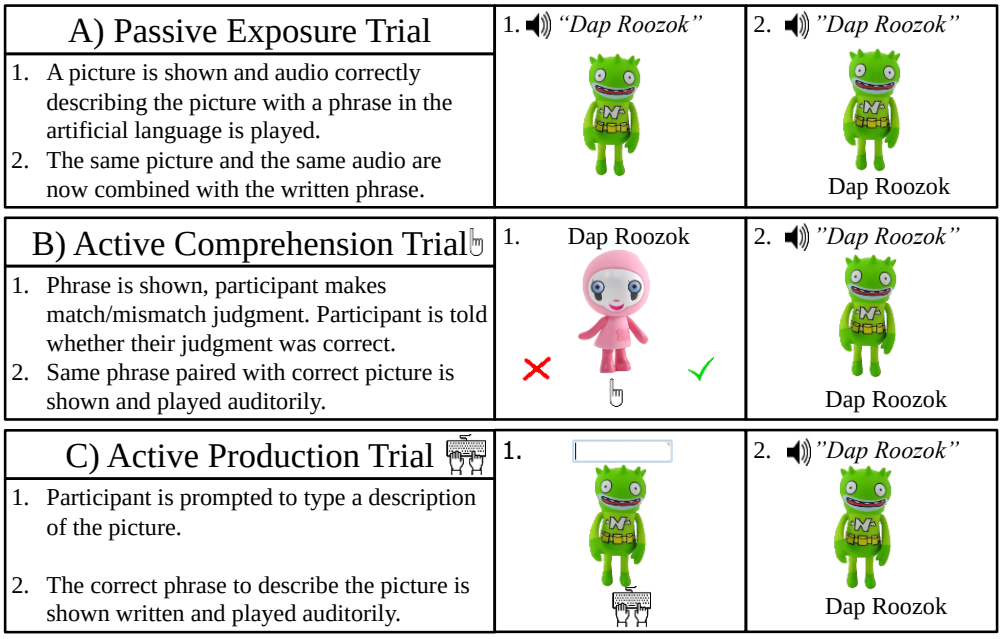
After this, at the end of training, there were four passive-only blocks of training to introduce the 12 unfamiliar aliens. Before I introduced these passive-only exposure blocks at the end of training, pilot participant performance on several tests was at floor for unfamiliar aliens. In these blocks, each unfamiliar alien featured once as singular and once as plural in a passive exposure trial. Note that the unfamiliar aliens were only seen in 2 passive trials, whereas the familiar aliens were seen in 4 passive and, importantly, 4 active training trials, thus maintaining a large familiarity difference. I reasoned that two passive exposure trials was little enough exposure that the unfamiliar aliens could still be considered as novel lexical items to test generalization performance on, particularly because participants did not do active comprehension- or production training trials with these unfamiliar aliens.

Training Tasks. Participants in all conditions got to know the language during passive exposure trials, which were identical for all conditions. During a passive exposure trial, participants saw a picture and heard a phrase that described this picture in the artificial language (Figure 4A). Then, the same picture was shown, this time with the written phrase below it, and the same phrase was played, in order for participants to have ample opportunity to learn the new

phrase. Note that written phrases were not present in Hopman & MacDonald's (2018) original passive exposure trials, though they were used successfully in a different study comparing production and comprehension training that included a written production test (Keppenne, Hopman & Jackson, 2021). Participants had no explicit, active task during these trials, they were instructed to pay attention and try to learn the language. In order to contrast production and comprehension training in a balanced way, I designed two active training tasks.

Figure 4

Example Trials for all Three Training Tasks



Note. A) Example passive exposure trial – these trials were the same for participants in all conditions. B) An active comprehension trial. C) An active production trial.

Active Comprehension Task. During the first phase of an active comprehension trial, participants saw a picture on the screen and heard a phrase in the artificial language (Figure 4B). Participants indicated by button-press whether they believed the phrase matched or mismatched the picture. Participants then got feedback at the bottom of the screen to indicate whether their

match/mismatch judgment was correct. During the second phase of the active comprehension trial, participants heard the same phrase, this time accompanied by the correct picture and the written phrase. Thus, on match trials, the picture didn't change between the two phases of the active comprehension trial, but on mismatch trials the picture did change.

Active Production Task. During the first phase of an active production trial, participants saw a picture on the screen accompanied by a typing box (Figure 4C). Participants had to type the description of the picture in the artificial language. They indicated by button-press when they were done writing. During the second phase of the active production trial, participants saw the same picture, heard the correct description in the artificial language and saw the correct phrase written on the screen. Thus, the second phase of active production trials was exactly the same as the second phase of active comprehension trials.

I chose a written production test for this study for several reasons. While I can test lab-participants in a sound-attenuated room with a professional quality microphone, due to the COVID-19 pandemic participants completed this experiment at home using their own equipment, with varying microphone quality. Written responses are easier to collect and not dependent on individual participants' hardware quality. Finally, written production attempts are easier to analyze, since they don't require transcription and can simply be processed by a script computing levenshtein distance between the response and the target phrase.

Controlling for production and comprehension differences. The two active tasks I employed are balanced for a number of known differences between production and comprehension: reading experience, attention and making task-relevant choices (Carter & Ste-Marie, 2017; MacLeod & Bodner, 2017; Sanli et al., 2013). The tasks are roughly equated for reading experience: in the first phase of a comprehension trial, participants read a phrase that

may or may not be a correct description of the picture; in the first phase of a production trial, participants read their own written attempt, which may or may not have been correct. Furthermore, both tasks involved making task-relevant choices, thus ensuring that participants need to pay attention.

Testing

After training, participants completed 4 different tests: a 2 alternative forced choice test, a 4 alternative forced choice test (containing both items testing stem vocabulary and items testing grammar understanding), an error monitoring test and a production test.

Two Alternative Forced Choice (2AFC) Test. In a 2AFC trial, a participant saw two pictures on either side of the screen and heard a phrase in the artificial language (Figure 5; Hopman & MacDonald, 2018). Their task was to indicate by button-press, as fast as possible, which of the two pictures the phrase described. Dependent variables measured in this test were accuracy and Reaction Time (RT). RT was simply measured from the start of the first word, since in all trials that first word, the determiner, already carried the grammatical information needed to disambiguate which picture matched the phrase.

This short test assessed participants on their understanding of the grammatical regularities for singular/plural and the neighborhood categorizations on the unfamiliar aliens. There were two rounds, and each of the 12 unfamiliar alien noun phrases was tested once in each round, counterbalanced for singular/plural between the rounds and neighborhoods. The trials in a round were presented in random order.

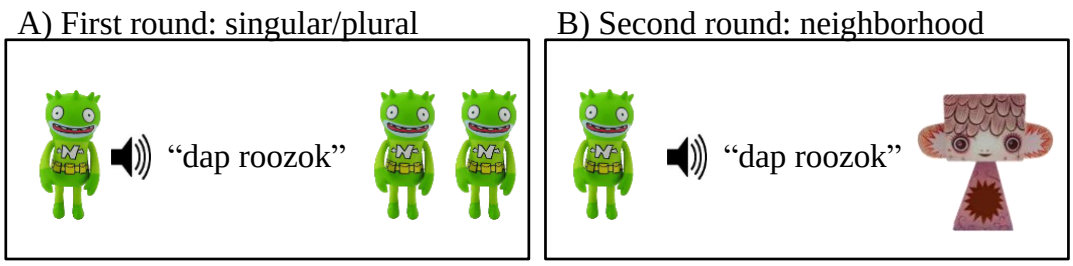
The first round tested participants on their understanding of the singular/plural rule (Figure 5A) by exposing them to a noun-phrase and showing them the same unfamiliar alien both singular and plural. If a participant learned which determiners and suffixes marked singular

and plural, they should be able to select the correct answer even if the lexical item itself and its visual referent were unfamiliar, by generalizing that the same determiners and suffixes could mark singular and plural even for unfamiliar lexical items. Specifically, these items assessed ability to generalize the singular/plural regularity within each neighborhood.

The second round tested participants on their understanding of the neighborhood categorizations and the grammatical markers for each neighborhood (Figure 5B). It did this by exposing participants to a target unfamiliar alien and its auditory description and, as a distractor, another unfamiliar alien *from the other neighborhood.* If a participant learned which visual features as well as which determiners and suffixes marked the large and small neighborhoods, they should be able to select the correct answer even if they had not memorized the specific lexical item – visual referent combination. Thus, these items assessed ability to generalize the between-neighborhood regularity. Finally, these items could show overgeneralization errors: participants might show a higher tendency to pick large neighborhood distractors for small neighborhood targets than the other way around.

Figure 5

Example Two Alternative Forced Choice (2AFC) Trials

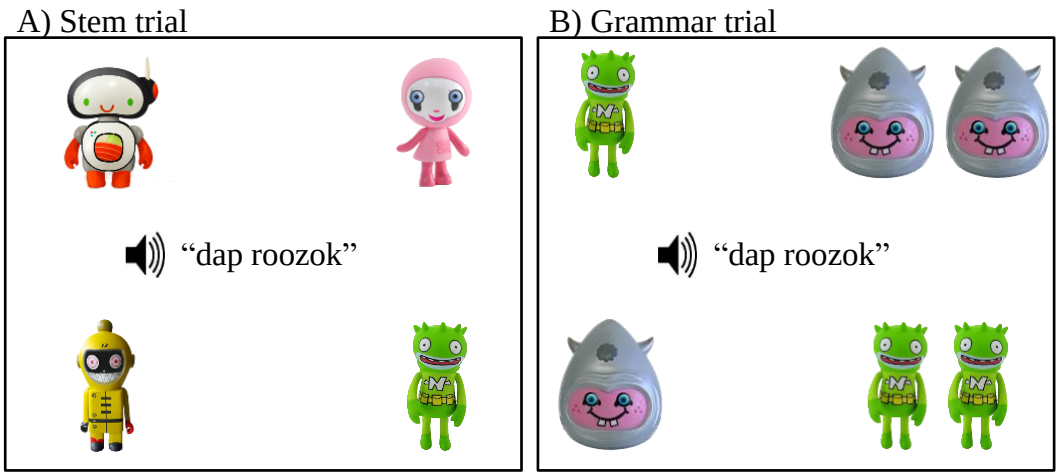


Note. Participant could use the A) plurality information or B) neighborhood information in the grammatical markers (determiner and suffix) to determine which of the two pictures corresponded to the phrase.

Four Alternative Forced Choice (4AFC) Comprehension Test. This test used a similar format to the 2 alternative forced choice test: the participant heard an auditory phrase and indicated by button-press which of 4 images matched the phrase. In this test, I wanted to assess learning of the stems as well as learning of the grammatical rules. Each of the 30 aliens (18 familiar aliens + 12 unfamiliar aliens) was tested twice, once in a grammar-targeting trial and once in a stem-targeting trial. If an alien's stem trial featured the singular noun phrase, its grammar trial featured the plural noun phrase (counterbalanced). Trials in this test were presented in a random order.

Figure 6

Example Four Alternative Forced Choice (4AFC) Trials



Note. A) In stem trials, participants could only use the noun stem to choose the correct picture, since all pictures were from the same neighborhood and had the same plurality. B) In grammar trials, participants could use grammatical information to choose the correct picture, since the distractors differed in plurality and/or neighborhood. Participants could use the noun stem instead of grammatical neighborhood information to choose the correct alien, but not the correct number of aliens.

Stem Trials. In a stem trial, participants heard a phrase and saw 4 different aliens on the screen, all from the same sub-neighborhood (familiar/unfamiliar and large/small) and equal in number (Figure 6A). Thus, the determiner and suffix could not be used to select the correct answer, and a participant needed to know the meaning of the noun stem in order to select the correct picture. These items assessed learning of the stems rather than the regularities. Since the first word to disambiguate the target in these trials was the noun, RT was measured from the start of the noun. Note that this could lead to a negative RT if a participant made a choice before the start of the noun.

Grammar Trials. In a grammar trial, participants heard a phrase and saw 2 different aliens (the target and a distractor alien from the other neighborhood) on the screen in both singular and plural providing 4 choices total (Figure 6B). Unfamiliar alien distractors were unfamiliar aliens from the other neighborhood, and familiar alien distractors were familiar aliens from the other neighborhood. There were more aliens trained in the large than in the small neighborhood. This meant that all of the 6 familiar small neighborhood aliens served as a distractor item twice, and that a random half of the familiar large neighborhood aliens never served as a distractor. With these constraints in place, familiarity of individual aliens could not be used to infer the correct answer (and thus could not drive overgeneralization behavior). RT was simply measured from the start of the first word, since in these trials that first word, the determiner, already carried the grammatical information needed to disambiguate which picture matched the phrase.

Recognizing the plurality and neighborhood of the determiner and suffix was enough to get these trials correct in principle (though the noun stem instead of the neighborhood rules could be used to select the correct alien but not the correct plurality). Trials with targets from the small

neighborhood and distractors from the large neighborhood were of particular interest since they provided an opportunity for overgeneralization (it is less likely that people would overgeneralize in the other direction and pick a small neighborhood alien when the target phrase described a large neighborhood alien).

Error Monitoring Comprehension Test. In an Error Monitoring trial, a participant heard a phrase in the artificial language, and saw the target alien depicted (cf. Keppenne, Hopman & Jackson, 2021). Pilot participant performance on this test was at floor when I initially did not include these pictures. This was likely due to participants being less focused while participating in studies from their own spaces. Note that the picture was never necessary to detect the grammatical errors, but might have made it easier to do so because neighborhood and plurality information were visually present in the picture of the alien. The participant's task was to indicate by button-press, as fast as possible, whether the sentence they heard was grammatical or not. Correct noun phrases were mixed with 3 different types of grammatical errors (see Table 2). Dependent variables measured in this test were accuracy and RT. RT was measured precisely, from the start of the first word in the phrase that could disambiguate whether the sentence was grammatical or not; for error types 1 (wrong plurality determiner) and 2 (wrong neighborhood determiner) this was the determiner, for error type 3 (wrong neighborhood suffix) and grammatically correct phrases this was the noun.

Since most cells of the 2x2 design consisted of six aliens (see Figure 3; all but the large, familiar sub-neighborhood), I wanted to test each of these six aliens on all these four different trial types to have as much item-based power as possible. For all six-alien sub-neighborhoods (all but the large, familiar sub-neighborhood, which consisted of 12 aliens), each alien occured

once in each of three different error types, and once in a grammatically correct phrase. This generated 72 trials (54 error : 18 correct), namely 24 per cell (6 aliens*4 trials each).

Then, to get the same item-based power, I also included 24 equivalent trials with aliens from the 12-alien large, familiar sub-neighborhood (six correct ones and six for each of the three different error types). However, this total of 96 trials was too unbalanced, with 3:1 error:correct trials. Thus, I added 16 more correct trials, all testing familiar, large neighborhood aliens. The resulting 40 trials using the 12 familiar large neighborhood aliens were chosen so that each alien occured at least three times total in this test (at least once correctly and at least once with an error).

Thus, the error monitoring test consisted of 112 trials (72 errors:40 correct; see e.g. Hopman & MacDonald, 2018 for a similar proportion of correct:error trials).Trials were counterbalanced for number within each sub-neighborhood as well as within each error type as well as per block. Within each block, the 28 trials were presented in randomized order, and block order was also randomized.

Table 2

Error Monitoring Example Phrases for all Trial Types for Both Neighborhoods

neigh.	plur.	correct phrase	1: wrong plurality determiner	2: wrong neigh. determiner	3. wrong neigh. suffix
large	sg.	**dap** rooz**ok**	**dap** rooz**ool**	**ked** rooz**ok**	**dap** rooz**esh**
	pl.	**lom** rooz**ool**	**lom** rooz**ok**	**jeb** rooz**ool**	**lom** rooz**aaf**
small	sg.	**ked** bur**esh**	**ked** bur**aaf**	**dap** bur**esh**	**ked** bur**ok**
	pl.	**jeb** bur**aaf**	**jeb** bur**esh**	**lom** bur**aaf**	**jeb** bur**ool**

Note. neigh.: neighborhood; plur.: plurality; sg.: singular; pl.: plural.

Error Type 1: Wrong Plurality Determiner. In the 24 trials of this type, the determiner's plurality was a mismatch with the suffix and picture plurlality. If participants were aware of the plurality error on the determiner, they should be faster and more accurate at indicating that

phrases like these are ungrammatical. These trials were included to provide a look at participants' ability to generalize the singular-plural rules to the relatively new unfamiliar aliens.

Error Type 2: Wrong Neighborhood Determiner. In the 24 trials of this type, the determiner was from the wrong neighborhood but matched the suffix and picture in plurality. If participants were aware of the appropriate determiner for each noun-suffix combination (and more generally that neighborhood), they should be faster and more accurate at indicating that phrases like these were ungrammatical. If a participant accepted a large-neighborhood determiner with a small-neighborhood noun, this was an overgeneralization error.

Error Type 3: Wrong Neighborhood Suffix. In the 24 trials of this type, the suffix was from the wrong neighborhood but matched the determiner and picture in plurality. If participants were aware of the appropriate suffix for each determiner-noun combination (and more generally that neighborhood), they should be faster and more accurate at indicating that phrases like these were ungrammatical. Note that, like in the wrong neighborhood determiner errors (type 2), accepting a large neighborhood suffix on a small neighborhood noun constituted an overgeneralization error.

Correct Phrases & Expected Overall Performance. This test included 40 correct phrases. Also, note that there were two different ways in which generalization is tested in the EM task. As mentioned, all four different trial types tested generalization ability by including trials with the 12 unfamiliar aliens. However, as indicated, only wrong neighborhood determiner (type 2) and wrong neighborhood suffix (type 3) errors tested for overgeneralization errors. Errors that participants made on wrong plurality determiner (type 1) and correct phrases for the unfamiliar aliens would still indicate trouble generalizing, but would not be examples of overgeneralization errors.

Production Test. I also tested participants' ability to describe the aliens, because typically production tests are harder for learners and are thus where overgeneralization errors are seen (e.g. the wug test; Berko, 1958; Wonnacott et al., 2008). In this test, participants saw a picture of an alien (either singular or plural) and had to type the alien's description in the artificial language. Note that this was identical to the first part of an active production training trial (but, unlike active production training trials, the production attempt here was not followed by the correct phrase to serve as feedback). Every alien was tested twice, once singular and once plural, leading to a total of 60 production test trials. The 60 trials were presented to participants in random order.

It is of particular interest to compare how often participants applied large neighborhood determiners and/or suffixes to small neighborhood nouns versus the other way around. Participants should be more likely to over-apply the more frequently encountered regularity associated with the large neighborhood to small neighborhood nouns (a classic overgeneralization error) than the other way around.

Procedure

Participants completed the experiment using their own equipment in a space of their own choosing. After signing up for the experiment online, participants received a link to complete the study in their own browser. They were instructed to complete the experiment in one go, in a quiet space without distractions, and to reserve 82 minutes of undisturbed time for completing the 2.75 credit experiment. They could complete the experiment at any time before the deadline of the timeslot they had signed up for.

After the electronic consent form, participants filled out a brief pre-experiment survey asking about demographic and language background information (see Appendix C). The

experiment started with two simple check trials during which participants heard a simple English word (e.g. “uncle”, “friend”) and were asked to type out the word they heard. On the first of these trials, participants had the opportunity to replay the audio as often as they wanted in order to adjust their computer volume. On the second and every further check trial, participants heard the audio only once without the option to repeat it – these check trials served as attention checks. Further check trials were included after every 3 training blocks (4 check trials), after the 2AFC test (1 check trial), after each 15 4AFC test trials (4 check trials), after each 28 EM test trials (4 check trials) and after each 15 production test trials (4 check trials). Each check trial was followed by a ‘5 minute optional break’ screen, where participants could push a button when they were ready to continue. If participants did not press the button to continue the experiment within the 5 minute timer (visible on the screen) running out, the experiment stopped and participants were directed immediately to the end-of-experiment survey. In this case, they received credit for the time spent doing the experiment. After the production test and the final check trial (this one without the option for a break), participants were automatically redirected to fill out a brief post-experiment survey (see Appendix C for details).

Results

Data processing

Accuracy and Reaction Time (RT) data were analyzed with separate (generalized) linear mixed effects models for each test, with maximum random effects structure (Barr et al., 2013). Where necessary because of non-convergence or singularity, the steps outlined by Barr and colleagues were used to simplify the random effects structure to achieve convergence and non-singularity. Each analysis included main effects for learning condition (production versus comprehension; henceforth: condition), neighborhood size (large versus small; henceforth

neighborhood) and familiarity (familiar versus unfamiliar), all two-way interactions and the three-way interaction. All 3 main predictors were centered, with learning condition as -0.5 (comprehension) and 0.5 (production), neighborhood as -0.5 (large) and 0.5 (small) and familiarity as -0.5 (familiar) and 0.5 (unfamiliar). Note that the 2AFC test only included unfamiliar aliens, and thus only had the other two predictors and their two-way interaction. Error monitoring accuracy data were additionally analyzed with a signal detection analysis to see if there were differences in *d*' scores for participants in the two learning conditions. RT analyses only considered correctly answered trials. Trials outside a participant's $M \pm 3SD$, as well as trials with an RT < 0.2 seconds were excluded. The written responses in the active production training and production test trials were pre-processed with a script in several analysis steps, and ambiguous production attempts were further processed by human coders (see Appendix D).

A preregistration done after piloting the final version of this study is available in Appendix F. The main text only considers learning condition and (over)generalization results; other results are discussed in Appendix G, which also contains all tables with regression results. Upon publication in a peer-reviewed scientific journal at the latest, all de-identified accuracy, RT and production test data, as well as analyses scripts and de-identified pre- and post-experiment survey responses will be made available publicly on OSF.io.

Comprehension Tests

Forced Choice Tests.

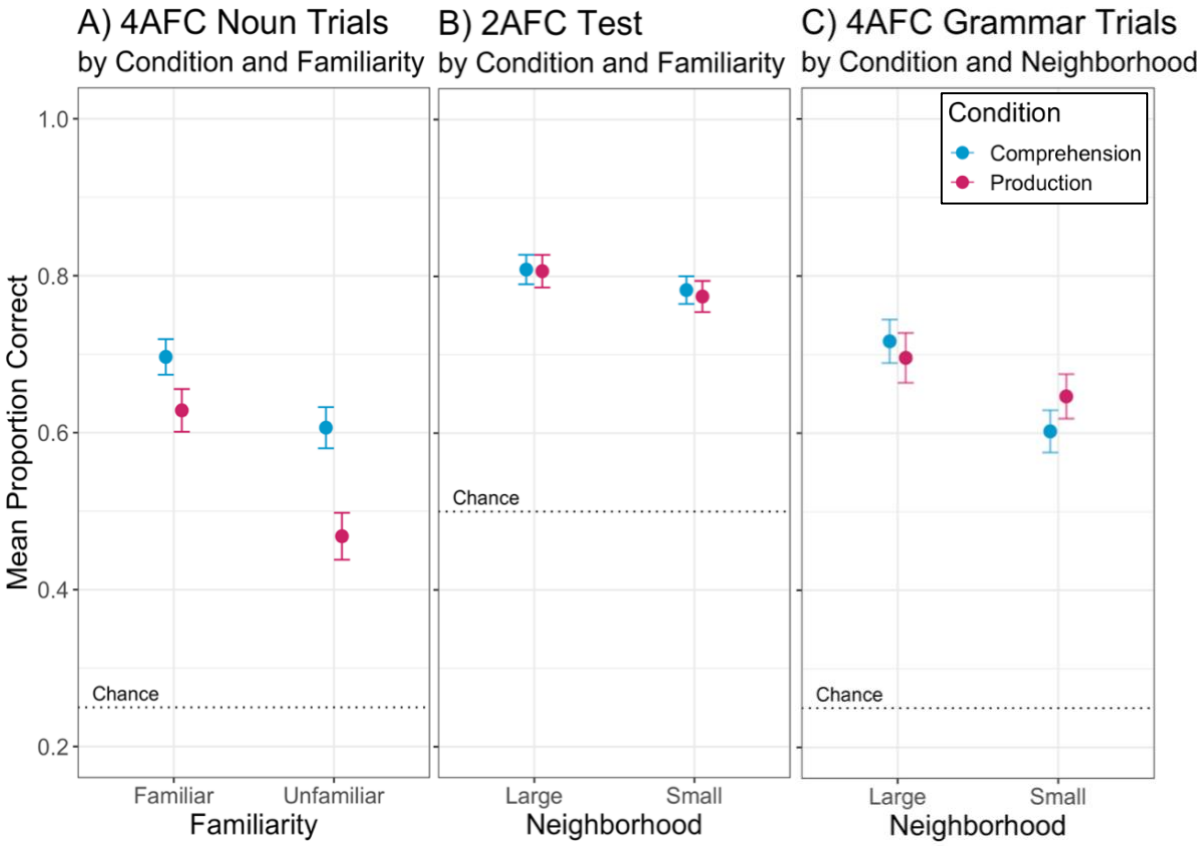
Stem Learning. To establish how well participants learned the stem vocabulary of the artificial language, 4AFC Noun Trials are reported first (note that these trails were part of the second test participants completed). These trials simply tested how well participants could map a stem (heard as part of a full noun phrase) to its corresponding alien (depicted amidst 3 distractor

aliens). Unexpectedly, participants in the comprehension learning condition were more accurate at stems than participants in the production learning condition (Figure 7A, Table G1), as shown by a significant negative main effect for learning condition. This advantage for comprehension participants was larger for unfamiliar compared to familiar alien stems, indicated by a signficiant condition:familiarity interaction. So, comprehension participants especially knew the unfamiliar alien-stem mappings better than production participants. Both of these results can be seen in Figure 7A (leftmost panel): the blue dots for comprehension participants are higher than the pink dots for production participants, indicating higher accuracy on stem learning for comprehension participants. The condition difference is larger for the unfamiliar aliens on the right than for the familiar aliens on the left. Comprehension participants were also somewhat faster than production participants on large neighborhood aliens, as shown by a marginal condition:neighborhood interaction (Table G2, Figure G2). All of these results indicated that comprehension participants learned the stem vocabulary better than production participants.

I had planned, and preregistered, to use a participant's overall performance on these stem trials trials as a covariate in all other analyses in order to control for overall vocabulary learning differences when analyzing the learning of grammatical regularities (as in Hopman & MacDonald, 2018). However, in planning this, I was expecting performance on these trials to not be significantly different between the two learning conditions. To prevent occluding condition differences elsewhere, I decided not to include this planned covariate in my other analyses. This also greatly improved model run time and convergence.

Figure 7

Main Forced Choice Comprehension Results



Note. Dots are model predictions, with error bars 95% Confidence Intervals (CIs).

Grammar Learning. Immediately after training, participants completed the Two Alternative Forced Choice (2AFC) Test targeting grammatical understanding of both the plurality and neighborhood information encoded in the determiners and suffixes. All items in this test were unfamiliar, and had been introduced in passive trials at the end of training, right before this test. Participants in both conditions successfully generalized in this test: both comprehension and production participants were signficantly above chance on this test (Figure 7B, Table G3). This is the first evidence that both comprehension and production participants successfully applied the learned grammatical regularities to novel lexical items (the unfamiliar aliens). However, participants also overgeneralized: overall, participants were significantly more

accurate on large than small neighborhood aliens. This is visible in Figure 7B (middle panel): performance is higher on the left for large neighborhood aliens than on the right for small neighborhood aliens. So, this first grammar test established that participants successfully generalized and applied the grammatical regularities to the novel, unfamiliar aliens. It also established that participants found it easier to apply the learned grammatical regularities to the more common large neighborhood, which is our first evidence that overgeneralization errors are happening. In this first grammar test, there were no condition differences.

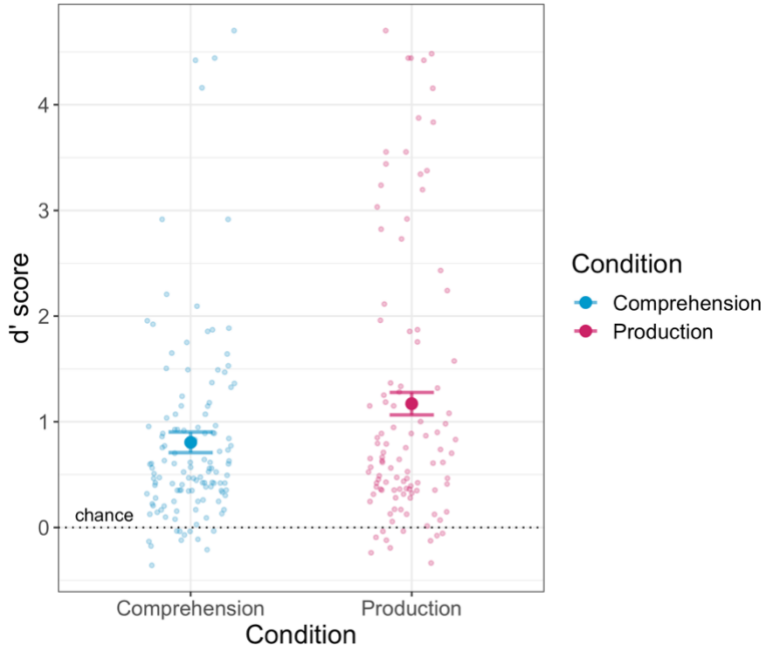
What about grammar trials in the 4AFC test? This test included both familiar and unfamiliar aliens, and that instead of 2, there were 4 choices on the screen. The 4AFC grammar items showed evidence of overgeneralizations, similar to the 2AFC test: participants were overall more accurate (Figure 7C, Table G6) and faster (Table G7, Figure G5) on large than small neighborhood items. This is visible in Figure 7C (rightmost panel): accuracy was higher for large neighborhood aliens on the left than for small neighborhood accuracy on the right. So, this is more evidence of overgeneralizations! Of course, the key research question concerns learning condition. There was no main effect of condition, but a significant condition by neighborhood interaction indicated that comprehension participants overgeneralized more than production participants: comprehension participants made relatively fewer errors on the more common large neighborhood aliens, and made relatively more errors on the less common small neighborhood aliens. In Figure 7C (rightmost panel), the blue comprehension and pink production dots swap for large versus small aliens. This is the first condition difference for overgeneralization errors in this study, and as expected, comprehension participants made more overgeneralization errors (small neighborhood errors) than production participants.

Error Monitoring Test

Main Analyses. Overall, production participants (M = 1.2, SD = 1.3) had significantly higher d' scores than comprehension participants (M = 0.8, SD = 0.9), showing that across all phrase types, production participants were better at discriminating between grammatically correct and incorrect phrases $t(232) = 2.54$, $p < 0.05$ (Figure 8). Separate models showed that for each of the three different types of grammatical errors production participants were significantly more accurate than comprehension participants (Tables G9, G11, G13, Figures G6, G8, G10). Since these were main effects, this means that production participants outperformed comprehension participants across the board, including on unfamiliar aliens testing generalization. This is the first condition difference for generalization in this study, and as expected, production participants generalize more accurately than comprehension participants.

Figure 8

d' Scores in the EM test by Condition

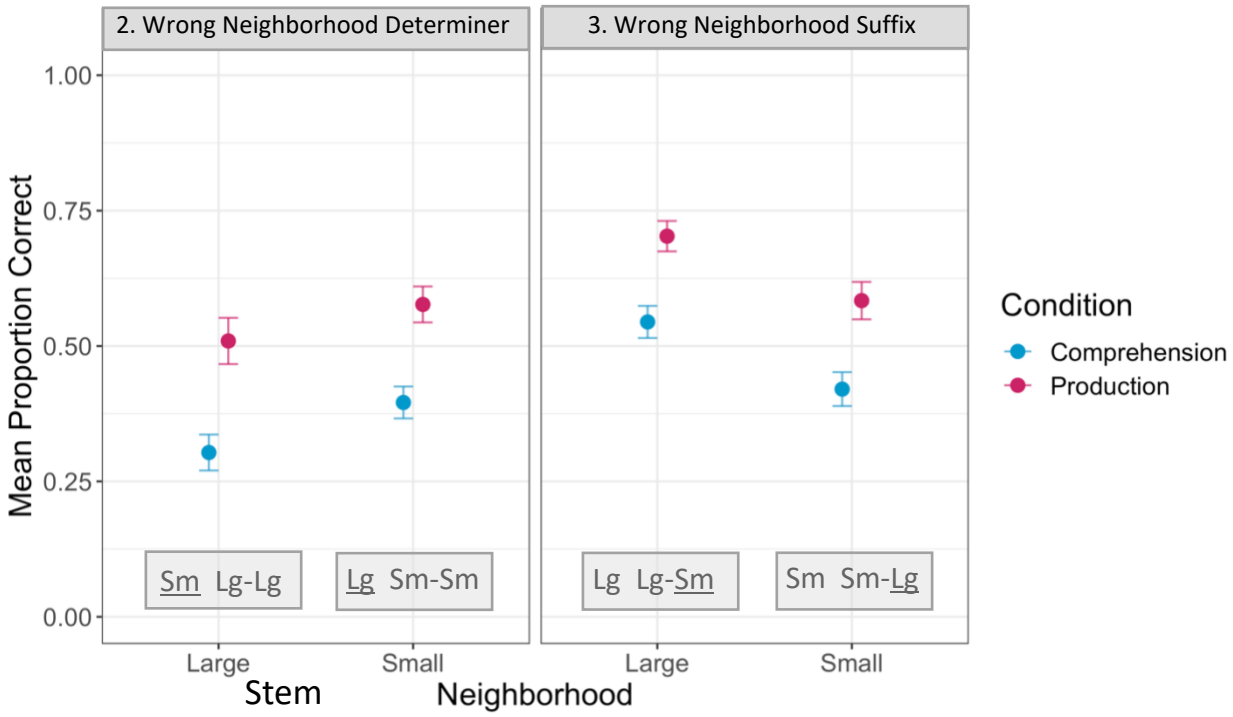


Note. Small dots are individual participant scores. Large dots are model predictions, error bars are 95% CIs.

Post-hoc Analyses. On the wrong neighborhood trials, in which either the determiner or the suffix did not match stem neighborhood, there was a potential for overgeneralization errors. Specifically, accepting a large neighborhood grammatical morpheme (type 2: determiner; type 3: suffix) on a small neighborhood trial (with a small neighborhood stem) would constitute an overgeneralization error. Instead, a serial order pattern emerged: when a large neighborhood morpheme was followed by a small neighborhood morpheme, participants were likely to catch the error (data labeled Lg Sm-Sm and Lg Lg-Sm in Figure 9). Conversely, when a small neighborhood morpheme was followed by a large neighborhood morpheme, participants were less likely to catch the error (data labeled Sm Lg-Lg and Sm Sm-Lg in Figure 9).

Figure 9.

Wrong Neighborhood Serial Order Effect (Stem Neighborhood by Error Location Interaction)



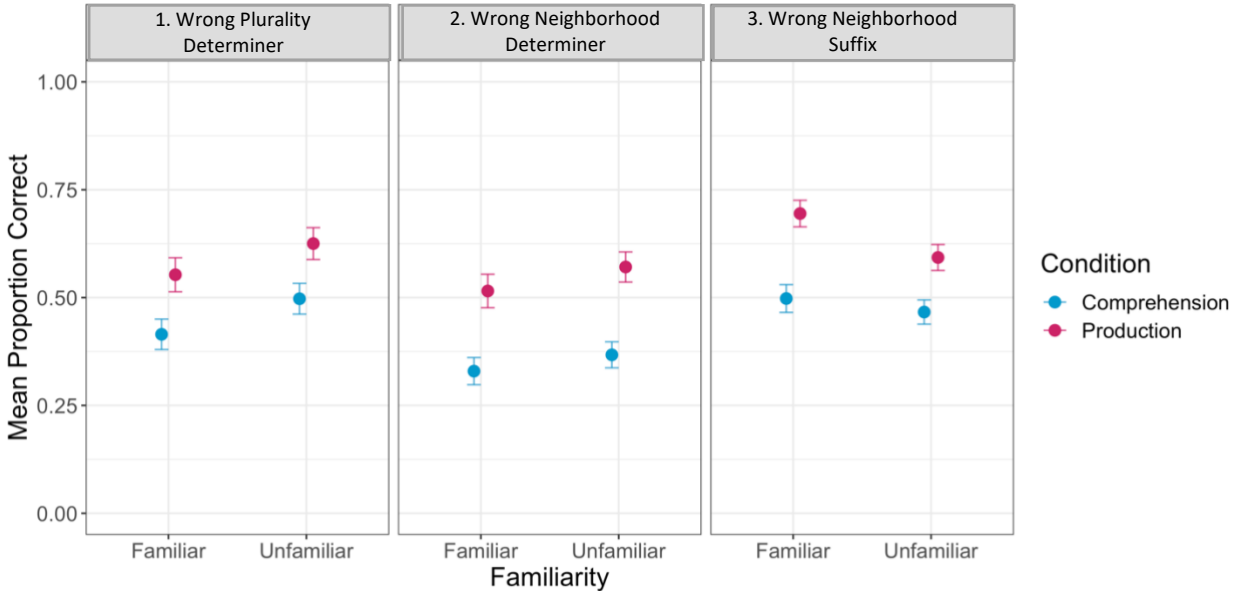
Note. Labels in text boxes illustrate the serial order effect (Sm = Small, Lg = Large; Determiner Stem-Suffix; error underlined). Dots are model predictions, error bars are 95% CIs.

A post-hoc, exploratory analysis confirmed that this stem neighborhood (stem part of the large versus small neighborhood) by error location (wrong neighborhood determiner versus suffix) interaction was significant (Figure 9, Table G15). Thus, large neighborhood morphemes followed by small neighborhood morphemes were easier to catch than the reverse serial order.

Another unexpected set of results merited a follow-up analysis. There were significant main effects for familiarity in all three error types, and this main effect of familiarity was in the expected direction only for wrong neighborhood suffix trials: participants did better at catching these errors on familiar than unfamiliar aliens. However, for both wrong plurality determiner and wrong neighborhood determiner errors, this main effect of familiarity was in the unexpected direction: participants did better at catching these errors for unfamiliar than familiar aliens. A post-hoc exploratory contrast analysis confirmed that participants were signficantly more accurate at catching determiner errors for unfamiliar aliens, and suffix errors for familiar aliens (Figure 10, Table G16).

Figure 10

Model Predictions for Error Location by Familiarity Contrast Analysis (see also Table G16).



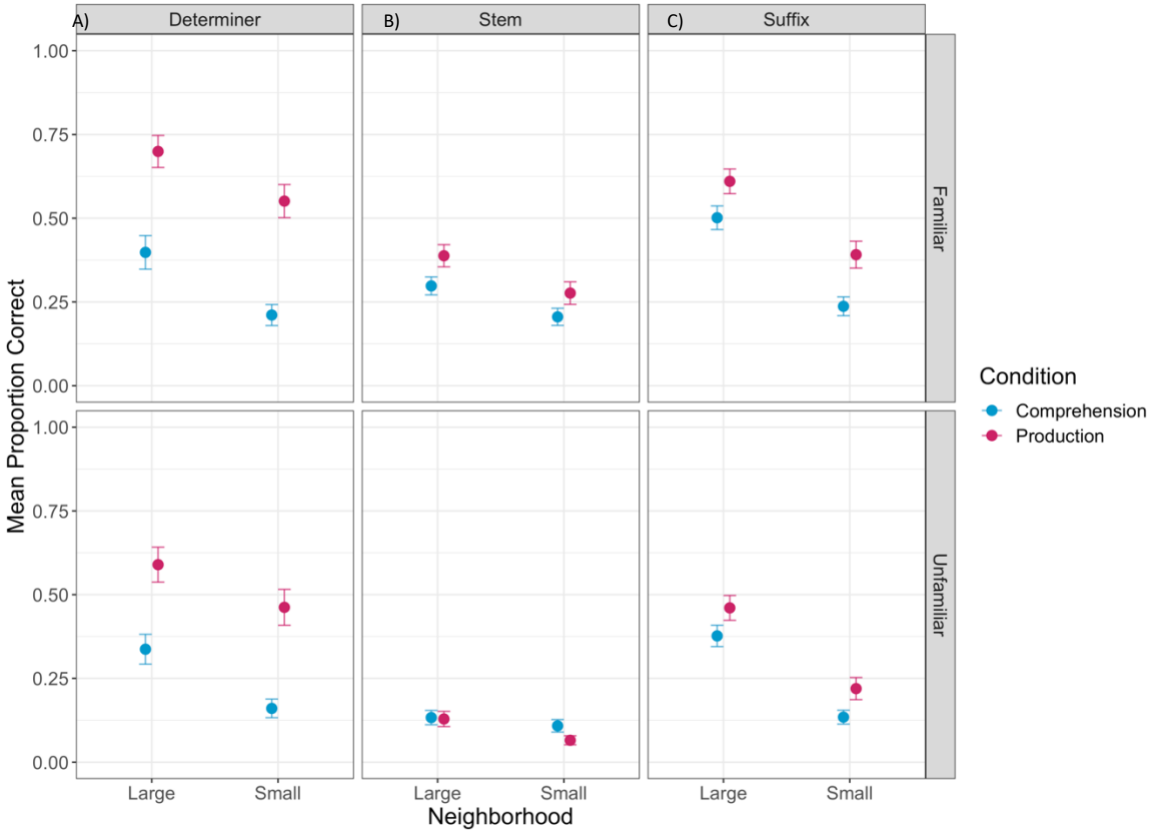
Note. Dots are model predictions, error bars are 95% CIs.

Production Test

Overall Accuracy. Overall production accuracy was analyzed separately for each morpheme. Not surprisingly, production participants significantly outperformed comprehension participants on both determiners and suffixes (Tables G17 & G19). This can be seen in Figures 11 A & C (the leftmost and rightmost panels): the pink dots for the production condition are overall higher than the blue dots for the comprehension condition. Production participants outperformed comprehension participants for grammatical markers overall, including for the on unfamiliar aliens testing generalization. Thus, production participants were better at generalizing grammatical regularities than comprehension participants.

Figure 11.

Production Test Overall Accuracy Results Per Morpheme



Note. Dots are model predictions, error bars are 95% CIs.

By contrast, for stems there was no main effect of condition (Figure 11B, Table G18). However, significant interactions of condition with neighborhood and of condition with familiarity show that there was a larger condition difference in favor of production for large compared to small neighborhood aliens, and for familiar compared to unfamiliar aliens. Interestingly, these interactions were further qualified by a marginal three-way condition by neighborhood by familiarity interaction, so that the production advantage was flipped in favor of comprehension for unfamiliar small neighborhood aliens. Thus, production participants were sometimes better, and sometimes worse than comprehension participants at producing stems. This can be seen in Figure 11B (the middle panels): the pink dots for the production condition are sometimes higher than, and sometimes lower than the blue dots for the comprehension condition. Thus, whereas production participants were better at learning and generalizing grammatical morphemes than comprehension participants, they were not unequivocally better at learning the stems.

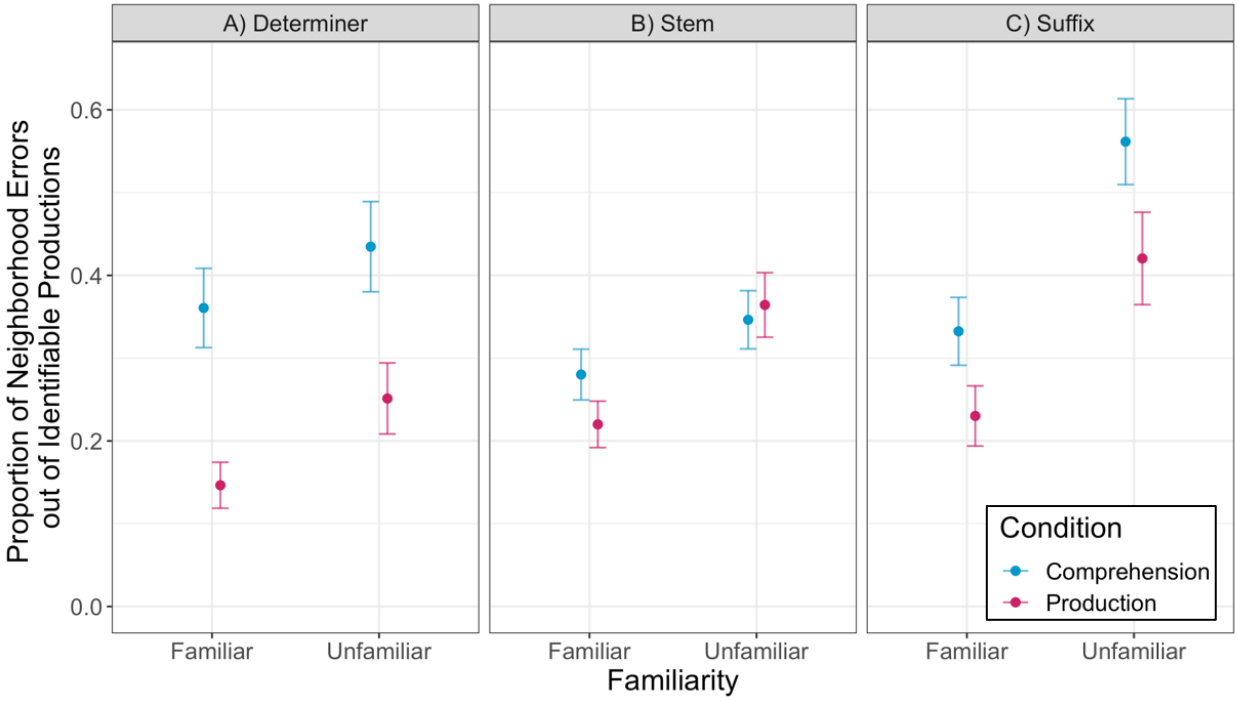
Overgeneralization. In order to assess overgeneralization, I analyzed, for each morpheme separately, what proportion of identifiable artificial language morpheme productions was part of the incorrect neighborhood. Thus, in this analysis the data were restricted to only trials on which a participant produced a recognizable artificial language form for that morpheme. Instead of analyzing correct/incorrect, as in the previous analysis, I analyzed whether the produced morpheme was a neighborhood error or not. Specifically of interest here were neighborhood errors on small neighborhood trials, because they constitute overgeneralizations: instances where participants produced large neighborhood morphemes for small neighborhood targets. Thus, the results and figure presented in the main text are the simple effects for the small neighborhood, obtained by centering the neighborhood predictor on the small neighborhood; the

model with the normal centered predictor for neighborhood is also presented in Appendix G, as is a figure that includes the large neighborhood (Tables G20-22, Figure G12). Having established these minutiae, I now turn to the actual overgeneralization results.

First, as expected, for all three morphemes, there was a main effect of neighborhood. The proportion of neighborhood errors was always larger for small than for large neighborhood aliens (Tables G23-25). Thus, as expected, the production test elicited many overgeneralization errors. Second, there was also a main effect of familiarity for all three morphemes, with the proportion of neighborhood errors always larger for familiar than for unfamiliar aliens (Figure 12). This is as expected: participants made more overgeneralization errors for unfamiliar than for familiar aliens.

Figure 12.

Overgeneralizations in the Production Test.



Note. Models depicted here present simple effects for the small neighborhood. Dots are model predictions, error bars are 95% CIs.

The key question was whether there were any differences in the amount of overgeneralization errors made by comprehension and production participants. There was a significant main effect of condition for determiners and suffixes (both grammatical morphemes; Figure 12A & 12C). Comprehension participants made significantly more neighborhood errors than Production participants on determiners and suffixes. For suffixes, we can see this in Figure 12C (the rightmost panel): the blue dots for comprehension participants are higher than the pink dots for Production participants, indicating that comprehension participants made more overgeneralization errors.

For determiners, this main effect was qualified by a marginal condition by familiarity interaction, so that the condition difference (comprehension participants making more neighborhood errors than production participants) was marginally larger for familiar than unfamiliar aliens. It is slightly hard to see in Figure 12A (leftmost panel), but the difference between the blue dots for comprehension participants and the pink dots for production participants is larger for familiar aliens on the left than for unfamiliar aliens on the right. So, comprehension participants overgeneralized more determiners than production participants, particularly when describing trained, familiar aliens. Thus, for overgeneralization, like we saw earlier in the production test for generalization, there are clear condition effects for grammatical markers: production participants generalize more accurately, and make fewer overgeneralization errors than comprehension participants.

What about overgeneralized stems? An overgeneralized stem is a trial where a participant used a large neighborhood stem to describe a small neighborhood depicted alien. For stems, there was no main effect of condition, meaning that there was no overall difference in how often

comprehension and production participants overgeneralized stems. However, there was a marginal condition by familiarity interaction, so that comprehension participants made marginally more overgeneralization errors on familiar stems, and production participants on unfamiliar stems. This is visible in Figure 12B (middle panel): for familiar aliens on the left, the blue dot for comprehension participants is higher, meaning that they overgeneralized more familiar stems. For unfamiliar aliens on the right, this is flipped, so production participants overgeneralized relatively more unfamiliar stems. So, like we saw earlier in the production test for generalization, there are no clear overall condition effects for stems: some stems are better generalized by production participants, other stems are better generalized by comprehension participants; production participants make fewer overgeneralization errors than comprehension participants on some stems, but not on other stems.

Discussion

I started this dissertation with the question whether production training, compared to comprehension training, would improve generalization of learned grammatical regularities to novel lexical items. In a literature review, besides identifying at best suggestive evidence that production *training* might indeed improve generalization, I also identified three patterns of overgeneralization results when production *tests* are contrasted with comprehension tests. These patterns were (a) more overgeneralization in production than comprehension tests, (b) different patterns of results in different comprehension or different production tests, and (c), a similar pattern of results in both production and comprehension tests. This combination of a research question about training with literature about testing led me to design an experiment in which I manipulated comprehension versus production *training,* and subsequently *tested* generalization

in a variety of different comprehension tests, as well as a production test. I found four major results.

First, across different comprehension and production tests, I found evidence that participants trained with active production trials (production-trained participants) outperformed participants trained with active comprehension trials (comprehension-trained participants) on generalizing a grammatical regularity to unfamiliar lexical items. Production-trained participants were better at applying the learned regularity to novel lexical items than comprehension-trained participants. Second, I found in several different tests that production-trained participants also made fewer overgeneralization errors than comprehension-trained participants. These results will be discussed in detail as they connect with the three patterns identified in the broader overgeneralization literature. Third, I unexpectedly found a benefit for comprehension participants on stem learning. Fourth, independent of learning condition, I found two serial order effects on error monitoring trials. All of these findings will be discussed in turn, after which I will also discuss implications for theories of statistical learning, limitations of the current experiment and applications and future directions based on these results.

Generalization

After learning the artificial language in this experiment, both sets of participants could successfully generalize the newly learned regularities, as shown by high performance on a two alternative forced choice test targeting grammatical understanding of unfamiliar lexical items. However, production-trained participants were better at catching grammatical errors in an error monitoring task and were better at producing morphemes that carry grammatical category meaning. Interestingly, these were all main effects, rather than interactions with familiarity. These results extend the learning benefit for production training found in Hopman & MacDonald

(2018) to generalizing to novel lexical items. However, it is not the case that production training carries a generalization benefit beyond the general learning benefit previously shown: the benefit of production training was roughly similar for applying grammatical regularities to learned and novel lexical items. Still, this answers my main research question with a resounding yes: compared to comprehension training, production training improves not just grammar learning but also generalizing.

So, production practice improves comprehension test performance, and even leads to better generalization of grammatical regularities on comprehension tests than comprehension training itself does. Conversely, even comprehension-trained participants who had never before produced the artificial language, produced over 25% correct morphemes when describing pictures with familiar aliens. Thus, my results clearly support shared representations between production- and comprehension processes, allowing for transfer (generalization across modalities). However, my results do not support full engagement of production-like mechanisms in comprehension processing, as e.g. Pickering and Garrod (2013) seem to suggest. Their account is admittedly underspecified enough that it's not exactly clear what they would predict to find in my experiment, but I interpret it as supposing fully shared representations and nearly lossless transfer between modalities. If that were the case, different training modalities would not lead to the differences in grammar learning and generalization that I find here and we found elsewhere (Hopman & MacDonald, 2018; Keppenne et al, 2021).

In Hopman & MacDonald (2018), we hypothesized that the binding of lexico-syntactic elements held in working memory during utterance planning in production is what improves learning during production, compared to the possibility of shallow, 'good-enough' processing during comprehension. However, it was unclear from those earlier results whether that

hypothesized binding was lexically specific, or whether production training would also lead to improved binding between, in this case, a set of visual features and a category of grammatical features, in absence of the specific (lexical) item. Here, I show that production-trained participants are better at catching grammatical errors, including on unfamiliar aliens, than comprehension participants. Production-trained participants also more accurately produced grammatical morphemes, including on unfamiliar aliens. Both of these are evidence of better generalization after production training than after comprehension training. These results are in line with the interpretation that the hypothesized binding between (abstract) categories, even if always happening during learning in the presence of specific lexical items, is strong enough to also persist beyond the context of those specific lexical items.

Overgeneralization

I also found overgeneralization errors in several tests. On the production test, for small neighborhood aliens, comprehension-trained participants made more neighborhood errors – overgeneralizations - on grammatical morphemes than production-trained participants. On the four alternative forced choice grammar items, all participants made more errors for small than large neighborhood target phrases, and a significant condition by neighborhood interaction showed that comprehension-trained participants were less accurate than production-trained participants on these small neighborhood items particularly. This second result is particularly important for my own research question about production versus comprehension training: I show that comprehension-trained participants make more overgeneralization errors, *even* on a comprehension test that is closer to their own training task. This shows convincingly that comprehension training leads to more overgeneralization errors than production training does.

At first glance, this main effect for condition in the production overgeneralization test, compared with a fairly subtle condition by neighborhood interaction in a comprehension test, conforms to overgeneralizaiton error pattern (a) from the literature: as expected, there is more evidence of overgeneralizations in the production test results than in the comprehension test results. However, like in the reviewed literature, a closer look confirms that the other two patterns of results are also present in my data. I found evidence of overgeneralization errors in one comprehension test, but not in the other two where I was expecting to also find overgeneralization errors: the two alternative forced choice test results were at ceiling, and the error monitoring test instead showed serial order effects. This is an example of pattern (b): different comprehension tests show different overgeneralization results. Finally, though the evidence for overgeneralization errors may have been stronger in the production than the comprehension test, both tests led to the same conclusions: comprehension-trained participants made more overgeneralization errors than production-trained participants. This is in line with pattern (c): a comprehension and a production test leading to the same conclusion about overgeneralizations.

Thus, interestingly, my experiment manipulating production and comprehension *training,* and using both production and comprehension *tests* to assess (over)generalization, confirms both that production *training* leads to better generalization (and comprehension training leads to more overgeneralization errors) and that all three patterns of overgeneralization results found in the literature contrasting comprehension and production *tests* can show up in a single experiment. While this means that the classic story of more overgeneralizations in production than comprehension tests holds, this also shows that test modality has other, more subtle ramifications for (over)generalization.

Just because one test, or one testing modality, does not show a certain expected pattern of overgeneralization results, this clearly doesn't mean another test wouldn't show the expected pattern of overgeneralization results. So a wiser approach, especially in resource-intensive multi-day learning experiments, might be to always implement several different tests after learning (e.g. Hudson Kam & Newport, 2005, 2009; Wonnacott, Boyd et al., 2012; Wonnacott, Newport et al., 2008). That way, if real effects are present, they are more likely to be detected, even when it is not clear a priori how well learners will do and which tests may not be sensitive due to floor or ceiling effects. This was part of the reasoning for my own experiment to include several tests of different levels of difficulty.

Finally, two patterns of results were present neither in the literature review nor in my own results. Due to task demands, it seems unlikely that a comprehension test would ever show more overgeneralizations than a production test. Furthermore, while there are plenty of examples both in my own data and the literature where one test or test modality shows results that another one doesn't, it also seems unlikely to find conflicting overgeneralization results in the same experiment. In practice, it may be unwieldy to conduct four separate tests after training, as this experiment did, especially when e.g. working with very young participants. Thus, if the main interest is finding overgeneralization errors, a production test would be the most likely to do so (Schwab et al., 2018). Whereas I was skeptical after my literature review, I now think that in certain cases, building theories and drawing conclusions from (over)generalization results in one modality seems warranted.

However, I would still argue that in general it is important to take modality into account. In my own experiment, since I contrasted comprehension and production modality in training, the more convincing result to show that comprehension participants make more

overgeneralization errors than production participants is the one from the comprehension test. Similarly, if a theory speaks particularly to comprehension mechanisms, like pre-emption as a way to retreat from overgeneralization errors (Goldberg, 2016; Tachihara & Goldberg, 2019), comprehension results are more convincing than production results, and if production is invoked in explanations, the relationship between comprehension and production during learning should be made more explicitly clear. Thus, the only situation in which building theories and drawing conclusions from one modality is warranted, is when that theory explicitly only involves that modality.

Vocabulary learning

Interestingly, I found that comprehension-trained participants outperformed production-trained participants on stem learning in this experiment. In prior experiments, we had trained participants to ceiling on vocabulary, and found no such differences. At first glance, the improved stem learning may seem contradictory to results in the memory learning literature, where retrieval-based learning has been shown to increase vocabulary learning compared to recognition practice (Karpicke & Roediger, 2008). However, the recognition-based control tasks used in retrieval-based learning experiments aren't typically balanced for task demands. In our experiment, comprehension training specifically also included making active, task-relevant choices. Thus, this discrepancy with retrieval-based learning results is reminiscent of how Second Language Acquisition experiments have found benefits for comprehension training, in experiments where meaningful comprehension tasks are contrasted with unbalanced production tasks that don't encourage in-depth grammatical processing. Similarly, in this case, recognition control conditions in retrieval-based learning experiments don't encourage in-depth memory processing and thus are found to be less effective for learning. In the only outright vocabulary

learning experiment that I myself have run (without grammatical regularities present), which used the same balanced production and comprehension training tasks, we found no condition differences in vocabulary learning (Hopman & Zettersten, 2018, and unpublished follow-up results). However, that experiment was designed to answer a research question about category learning, and yielded ceiling results for vocabulary learning. In order to understand the vocabulary findings in the present study better, it would be interesting to similarly conduct a vocabulary learning study with balanced comprehension and production training tasks, that doesn't train vocabulary to ceiling. That way, condition differences might become visible. Before running the present study, I would have predicted better vocabulary learning for production-trained participants in such a study, but I am now wondering whether comprehension-trained participants might actually be better at vocabulary learning.

The results of the experiment presented in this dissertation lend themselves to another thought-provoking interpretation. At least with the learning procedure used in this study, comprehension and production practice seem to differentially impact how learners process *novel* lexical items. Production-trained participants generalized the grammatical regularities better to these novel lexical items, whereas comprehension-trained participants learned the vocabulary (the stems) better. I view this as parallel to a category learning finding discussed in my literature review. When participants are asked to explain out loud during learning why exemplars belong to a category, this improves their ability to learn patterns, but hinders their ability to learn unique features of exemplars (Williams et al., 2013). Thus, production and production-like tasks may generally encourage and improve regularity learning, but may hinder item learning compared to comprehension and comprehension-like tasks, at least when the two modalities are contrasted in balanced ways.

Serial order and non-adjacent dependencies

In the error monitoring test, instead of evidence of overgeneralizations, I found a serial order result. Participants in both learning conditions were better at catching errors that involved small neighborhood morphemes following large neighborhood morphemes than the other way around. This could indicate more predictive processing for large than small neighborhood morphemes. After a large neighborhood morpheme, participants expected other large neighborhood morphemes, whereas participants' expectations may not have been as strong after small neighborhood morphemes. Generally, we know that learners are sensitive to serial order patterns like these (e.g. Aslin & Newport, 2012), and it stands to reason that these expectations are sharper for the more common large neighborhood.

Unexpectedly, participants in both learning conditions were better at catching wrong determiner errors for unfamiliar than for familiar aliens. At first glance, this is weird: why would participants ever do better on phrases containing novel than trained items? Admittedly, I am not sure why this effect appeared in my data, and before drawing strong conclusions from this effect it would be important to replicate it under other circumstances, e.g. by running a well-powered in-person version of this experiment. However, in the interest of future research and sparking ideas, I will speculate here on what may have caused this unexpected benefit on unfamiliar alien wrong determiner errors.

It is possible to perform perfectly on the error monitoring test in our experiment by simply knowing the four determiner and suffix pairs (one for each of the four neighborhood by plurality combinations). We know from previous research that adults can learn this type of non-adjacent (or AxB) dependency, and that variability in the intermediate word helps learning (Gómez, 2002; Onnis et al., 2004; Romberg & Saffran, 2013). Interestingly, in our data

participants were better at catching these non-adjacent AxB errors for unfamiliar middle 'x' elements particularly when the errors occured on the determiner, the 'A'. Conversely, when the non-adjacent AxB error occurs on the suffix, the 'B', participants were better at catching the errors with familiar middle 'x' elements. Thus, there is a serial order effect present in our non-adjacent dependency error detection. If the error occurred on the A element, the following familiar middle 'x' element seems to have distracted learners from the AxB regularity. If the error occured on the B element *following* a familiar middle 'x' element, participants were better at catching the error. I don't know of any similar results in other experiments concerning non-adjacent AxB learning. However, given that variability during learning is helpful in picking up on non-adjacent AxB dependencies, I speculate that unfamiliar middle 'x' elements could have allowed learners in my experiment to better focus on the non-adjacent AxB dependency.

Implications for theories

Looking back to my literature review, there are some areas and debates where my findings have ramifications. In the past tense debate, I see my results showing the impact that learning (and testing) modality can have on generalization, as fitting in with single-mechanism, connectionist views. Specifically, there is an implemented computational model that can simulate several different production and comprehension tasks (Woollams et al., 2009). Woollams and her co-authors did not explicitly test generalization, nor did they implement production versus comprehension as a training contrast, and it is thus an empirical question whether their model would show the production benefits that my data show. However, they did explicitly construct their model to be able to handle different naturalistic verb inflection tasks, and because of this, it is conceivable to test my findings. Interestingly, before that model was published, Pinker (2006, pp. 224) criticized single mechanism theories for being able to handle only production but not

comprehension of inflected forms. However, I have not been able to find any references to task modality in dual-mechanism, words and rules accounts (e.g. Pinker, 2006). However, if such an account were to want to explain production's role during learning and generalization, it would first have to come up with a model that implements both, or it would at least have to specify the role of modality explicitly.

Unfortunately, this lack of specificity with regards to production's role during learning is reflective of a broader criticism of the first language acquisition literature, as the following quote illustrates:

> In current theories of first language acquisition, children's productions are of interest primarily as evidence of what children know and not as a potential contributor to the development of language. Depending on the theoretical approach, language development is held to be the result of different contributions from innate structure, analytic abilities, and information provided in the speech children hear (Hoff, 2015). There is no theoretical claim that output does not matter; rather, it is just not given much attention as a potentially relevant factor. (Ribot et al., 2018, p. 929).

This lack of interest in language production as affecting learning is also evident in reviews on statistical learning. Aslin & Newport (2012) don't mention task modality, but implicitly seem to only focus on input and input processing, even though, like me, they define rule learning (a skill often assessed in production-like tasks) as the abilitiy to apply rules to novel instances, and argue that rule learning is an outcome of statistical learning mechanisms in the right contexts. I agree with Frost, Armstrong and Christiansen's (2019) remark that statistical learning research "often [is] too vague and imprecise regarding actual representations, processing mechanisms, and learning outcomes" (p. 1147). Those authors explicitly criticize a 'unitarian'

view of statistical learning, and state that "a good SL [Statistical Learning] theory is one which considers and focuses on the interactions of the organism with the environment" (p. 1139), and contrast this with a view of statistical learning in which the learner is a passive absorber of environmental distributions. It is striking then that in this review, the 'pluralistic' approach that the authors argue for is specified mainly as needing to take into account different *input* modalities and domains in which statistical learning has been researched (learning about e.g. faces, tones, syllables, shapes). My results also speak to the nature of regularity learning in general and generalization specifically. To the extent that my data show modality differences in learning and generalization, it is not enough for theories of regularity learning and generalization to remain modality-agnostic. Neither comprehension/input/perception-only, nor modality-agnostic views on generalizaiton can explain why production training would improve generalization compared to comprehension training.

MacDonald's (2013) Production-Distribution-Comprehension framework provides a compatible account for the important role that production plays at a larger scale. Production-pressures, like serial order and hierarchical processing during utterance planning, influence what people say. These utterances, shaped by individual, in-the-moment production pressures, when zoomed out together form language distributions. It is these distributions than comprehenders then pick up on. The review article quoted earlier also notes the dearth of mechanistic explanations of statistical learning phenomena (Frost, Armstrong & Christiansen, 2019, pp.1147). While I agree with that observation, I think a fruitful path to attaining a mechanistic view and more specific, testable theories needs to incorporate task modality more broadly, and language production more specifically. Language production is relatively well-understood at a mechanistic level, and scientists have long worked with explicit ideas of representations during

different processing stages (e.g. Levelt, 1993). Converesely, these well-specified, mechanistic views of language production do not typically encompass ideas about learning and change, but rather attempt only to capture the adult 'final state'. Updating those theories to include learning is beyond the scope of this dissertation, but it is clear that cross-polination of these ideas and theories would have benefits on both sides. I now turn to an example of how implementing specifics about task modality has helped a different area of theorizing become more mechanistic.

The 'production effect' in the memory literature is a well-replicated phenomenon where participants remember words in a list they said out loud better than words they heard another speaker say out loud. Whereas earlier work on this 'production effect' within the memory literature seemed focus on the comparative distinctiveness of memory traces, recent work has been more explicit about possible mechanisms and how production- or perception-learning interacts with the language system in general. Specifically, Kapnoula and Samuel (2022) show evidence that production might initially help during early word-learning, and later hinder, leading them to posit that production facilitates encoding but interferes with retrieval, whereas comprehension/perception might facilitate this integration. There is also evidence that, in novel sound learning, immediate production might hinder, interfering with encoding, whereas providing participants with a delay between hearing a sound contrast and trying to produce it mitigates this interference (Baese-Berk & Samuel, 2022). The generalization that is of interest in this line of work is the ability to perceive the studied sound contrast when produced by a novel speaker. Thus, while the 'production effect' is still often explored as a between-item, within-participant phenomenon, generalization is of interest in this type of work too.

There is an important difference between our work and these lines of work, that is reflected in both theoretical framing and practical implementation and helps reconcile the

different findings. Kapnoula and Samuel (2022) explicitly note retrieval-based learning as a nuisance variable, which they control for by having participants in their production conditions repeat a word, rather than generate it from memory. This way, they conceive of their task as only tapping into phonological short term memory, which is of interest to them (note that the idea of a separate phonological short term memory is itself in contrast with theories that view verbal working memory as an emergent property of language processing rather than a separate system, Schwering & MacDonald, 2020). In contrast, I explicitly theorize that it is the binding during utterance planning, when people have to retrieve morphemes and engage in hierarchical and serial order processing, that is responsible for the improved learning I see in my production condition. Thus, I do not see our results and theirs as contradictory, since we are explicitly tapping into broader production-comprehension differences. Interestingly, Baese-Berk and Samuel (2022) speculate that production might help with learning rule-based systems, e.g. the morphosyntactic learning we've tested, but not with less rule-based phenomena like learning to perceive a novel sound contrast. This latter supposition is very much in line with our result that, on novel lexical items, production participants do better at generalizing a learned regularity, whereas comprehension participants seem to do better at learning the new stems, which are less rule-based.

Limitations

A main difference between the results of this experiment and earlier published worked using a similar training paradigm (Hopman & MacDonald, 2018) is the (relative) lack of reaction time differences between conditions. Reaction time differences are of interest because fast, automatic processing of grammatical features is a difficult-to-achieve halmark of native-like language comprehension (Grüter et al., 2012). There are two main differences between the two

experiments that may explain the lack of robust reaction time differences in the current study. First, the current experiment was run online, with participants completing the experiment in their browser in their own space. While I did ask participants before starting the experiment to confirm that they were in a quiet space without distractions, many participants indicated that there had been some noise and/or distractions in their space in the post-experiment questionnaire. These distractions could, and did, lead to noisier data, as evidenced by the amount of unusable data I collected. Beyond affecting the overall quality of the data, participants also used different operating systems, browsers, hardware and internet connections to complete the experiment, and all of these can affect the precision of RT measurements. I also collected a pilot sample of in person participants (n = 39), and evaluated those data using the same usability criteria, and found data loss more comparable to earlier in person experiments, and markedly lower than for the online self-paced data collection done here. The pilot sample was not large enough to definitively conclude whether reaction time data would have shown condition effects in person, but the data did seem less noisy, even after applying usability criteria.

The second big difference was the language used in this study compared to the artificial language used in Hopman & MacDonald (2018). Specifically, the balance between length and complexity of phrases versus the number of stems learned was very different. Our 2018 study included only 18 ‘stems’ (vocabulary words of different grammatical categories), but those 18 different words combined to form 7-word-long full sentences. In contrast, this study had 30 unique stems, which combined with determiners and suffixes in limited ways to form noun phrases. This study employed 30 unique stems so that I could create a familiarity by neighborhood interaction and still have good item-level power, but it’s hard to say what effect this emphasis on stem-memorizing might have had on reaction times. It is also likely that

reaction time differences become larger for longer phrases. For example, in the Hopman & MacDonald's (2018) data, reaction time differences on the forced choice test, consisting of mostly 3-5 word phrases, were relatively small, and reaction time differences on the error monitoring test, consisting of 7 word full sentences, were larger. It is hard to say whether the shorter phrase length in this study otherwise affected results. Furthermore, it is not possible from these data alone to disentangle whether differences with earlier results were due to the noisy data or due to the shorter phrase length.

Finally, results from lab-based artificial language learning studies do not always replicate with classroom learners of natural languages (Paul & Grüter, 2016; but see Ettlinger, Morgan-Short, Faretta-Stutenberg & Wong, 2016). In order to address that limitation, Keppenne et al. (2021) implemented a version of Hopman & MacDonald's (2018) training paradigm to teach early L2 German learns about German grammatical gender agreement. That study both replicated and extended Hopman & MacDonald's artificial language findings. The production-trained group outperformed the comprehension-trained group on comprehension *and* production tests tapping into German grammatical gender agreement. Because of those findings, I am reasonably confident that a similar extension testing generalization of German grammatical gender agreement to novel lexical items would yield improved generalization for production-trained participants, as well as increased overgeneralization errors for comprehension-trained participants. What I am more curious about is what such a natural language version of the current experiment would find for vocabulary learning on both trained and novel lexical items. The stems in the artificial language used in this dissertation were designed to be pronouncable to the English native speaking participants I recruited. It is conceivable that the added complexity of learning words with non-native phonotactics would bring out even larger production versus

comprehension differences in vocabulary learning than the present study did. If so, results like those might dictate which studying strategy is more effective to use depending on whether vocabulary or grammar learning is more important in a given classroom or to a given learner.

Applications and Future Directions

The data presented here replicate prior research showing that, in a balanced design, language production training leads to better grammar learning than language comprehension training (Hopman & MacDonald, 2018; Keppenne et al., 2021). This study further extends that finding to show that production training also improves generalization of learned grammatical regularities to novel lexical items, both when tested in comprehension and in production tests. These findings corroborate our conclusions from earlier work that language production practice can improve grammar learning and even grammar comprehension, and is thus an effective tool for second language grammar instruction. In contrast to my current and earlier results, a meta-analysis in the field of second language acquisition comparing comprehension- with production-based instruction concluded that comprehension-based instruction results in better comprehension at immediate post-test. A crucial distinction is, again, how exactly comprehension and production are implemented. In a review of the second language acquisition literature, DeKeyser and Botana (2015) point to the "drill-like nature" (p. 301) of how production-based instruction is often implemented in this literature as a reason for comprehension-based instruction advantages. They noted that when production-based instruction is implemented in a balanced way with comprehension-based instruction, second language acquisition studies find an advantage for production-based instruction like I do here.

Thus, how exactly language production practice is implemented is crucial for second language instructors who wish to use language production as an effective tool for grammar

teaching. For improving grammar learning, it is key that production-based activities include meaning-based, generative production, that involves retrieving vocabulary and grammatical morphemes from long-term memory and engaging in the hierarchical and serial order processing inherent to utterance planning. While simple drill-like production exercises, like repeating a phrase or reading a provided text out loud might benefit other areas of second language acquisition like pronunciation, they do not provide the in-depth processing that benefits grammar learning. And it is of critical importance that, if second language acquisition researchers wish to compare the benefits of production-based versus comprehension-based instruction, they implement the two in balanced ways.

One particular aspect of the way in which I've implemented balanced comprehension versus production training merits follow-up research: the role of feedback. In order to allow for error-driven learning (see e.g. Clark, 2013 for human learning; Rumelhart et al., 1988 for computational modeling), and to prevent learners from persisting with their own early errors, active learning trials in my studies always include 'feedback' in the form of the correct audio-picture pairing after the learner has made their own match-mismatch judgment or production attempt. In (non-language) education contexts, feedback has been shown to improve learning from both comprehension-like multiple choice tests (Butler & Roediger, 2008) and production-like short answer tests (Kang, McDermott & Roediger, 2007). Notably, in the field of second language acquisition Swain's (2005) 'output hypothesis' also includes an important role for feedback and a learner's ability to notice discrepancies between their own (potentially ungrammatical) productions and native speakers' (grammatical) productions. As far as I'm aware, the role of feedback hasn't been directly tested in any experiments comparing the effects of *balanced* production and comprehension training for the purpose of grammar learning. This is

an important avenue for future research, because typical classroom situations, with many learners and one teacher, cannot always easily incorporate immediate, individual feedback for meaning-based, generative production tasks in the same way that a computer-based experiment can.

Another potential application of the results presented in this dissertation to real life second language learning situations is in testing and assessment. Based on both my literature review and my own data, it is clear that different types of tasks are more or less likely to elicit overgeneralization errors. If a teacher wishes to assess early learners' grasp of grammatical regularities, a relatively easy forced choice comprehension or grammaticality judgment task may suffice to differentiate students' mastery of a given regularity. Conversely, if the goal of a test is to assess mastery of a second language to the highest standards (e.g. C2 level in the Common European Framework of Reference for Languages, see Council of Europe, 2001), production tests may be more likely to differentiate highly proficient learners. For example, for English, a test in which learners are asked to produce a past tense inflection given a stem can elicit errors even in native speakers (Woollams et al., 2009). Finally, standardized language tests like the TOEFL iBT, the IELTS and the Duolingo English Test that are used by universities to help determine admissions for students from non-English speaking countries typically provide learners and institutions with an overall score as well as different sub-scores (e.g. IELTS & TOEFL iBT: listening, reading, writing and speaking; Duolingo English Test: literacy, comprehension, coversation, production). Presumably, proficiency with grammatical regularities is important for each of these sub-areas; I would argue that each of these sub-areas thus needs modality/sub-area specific trials testing grammatical regularities. Some of these standardized tests now employ computerized adaptive testing, in which item levels adapt to a testee's accuracy on earlier items, thus allowing for more efficient testing (McCarthy et al., 2021). An intriguing

possibility might be to not just adapt which items are presented, but also which testing modality is chosen.

Conclusions

I conducted an experiment to test whether language production training improves generalization of learned grammatical regularities compared to language comprehension training. I found that yes, language production training improves generalization, and comprehension-trained participants make more overgeneralization errors than production-trained participants. In the existing literature, statistical learning is often approached and written about as an amodal phenomenon, which limits my ability to draw conclusions about how my results fit with curent theories. I argue here that, since my results show that production and comprehension can differently affect learning of regularities, a theory of statistical learning and generalization is not complete if it is either modality-agnostic or explicitly comprehension, perception, or input-focused only. Thus, in a way, my data are a challenge to the field of statistical learning to become more specific about modality, mechanisms and representations involved in learning and generalization.

In contrasting production and comprehension *tests*, I found three patterns of results both in the published literature and in my own data: (a) people make more overgeneralization errors on production than comprehension tests, but (b) often different comprehension and production tests may themselves be differentially likely to elicit overgeneralizations, and (c) results from production and comprehension tests do tend to allow for similar conclusions about overgeneralization. Thus, while a single elicited production test may be the most likely to show overgeneralization errors, it might make more sense to prod overgeneralization in different tests

using different modalities. Doing so is particularly important if results are meant to inform theories that consider both modalities.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701.

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: word learning, morphology, and verb argument structure: Overgeneralization in child language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4* (1), 47–62.

Ambridge, B., Pine, J. M., Rowland, C. F., & Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*(1), 87-129.

Ambridge, B., Pine, J.M., & Rowland, C.F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition, 123* (2), 260-279.

Ansari, M. S. (2015). Speaking anxiety in ESL/EFL Classrooms: A holistic approach and practical study. *International Journal of Education Investigation, 2*(4), 38-46.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science, 21*(3), 170-176.

Baese-Berk, M. M., & Samuel, A.G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89*, 23-36.

Baese-Berk, M. M., & Samuel, A. G. (2022). Just give it time: Differential effects of disruption and delay on perceptual learning. *Attention, Perception, & Psychophysics*, *84*(3), 960-980.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28.

Barik, H. C., & Swain, M. (1978). Evaluation of a French immersion program: The Ottawa study through Grade five. *Canadian Journal of Behavioural Science*, *10*, 192–201.

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.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 553.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2-3), 150-177.

Berko, J., & Brown, R. (1960). Psycholinguistic research methods. *Handbook of research methods in child development*, 517-557.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*(2), 201-210.

Bixby, K. N. (2017). *Production effects on perception: How learning to produce sound changes auditory perception.* Unpublished doctoral dissertation, Department of Brain and Cognitive Sciences, University of Rochester, New York.

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, *143* (1), 295–311.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, *112* (44), 13531–13536.

Brysbaert, M. & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1* (1).

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition, 36*(3), 604-616.

Carter, M. J., & Ste-Marie, D. M. (2017). Not all choices are created equal: Task-relevant choices enhance motor learning compared to task-irrelevant choices. *Psychonomic bulletin & review*, *24*(6), 1879-1888.

Choi, D., Bruderer, A. G., & Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of Bruderer et al. (2015). *Psychonomic bulletin & review*, 1-12.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2019). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important?. *Bilingualism: Language and Cognition*, 1-15.

DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2011). Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behavior and Development, 34* (4), 590–601.

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review, 30*(1), 215-228.

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. (2016). The relationship between artificial and second language learning. *Cognitive science, 40*(4), 822-847.

Eysenck, M. W. (1985). Anxiety and cognitive-task performance. *Personality and Individual differences*, *6*(5), 579-586.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive science*, *42*(8), 2670-2698.

Fortune, T. W. (2012). What the research says about immersion. *Chinese language learning in the early grades: A handbook of resources and best practices for Mandarin immersion*, *1*, 9-13.

Fung, Y. M., & Min, Y. L. (2016). Effects of board game on speaking ability of low-proficiency ESL learners. *International Journal of Applied Linguistics and English Literature, 5*(3), 261-271.

Goldberg, A. E. (2016). Partial productivity of linguistic constructions: Dynamic categorization and statistical preemption. *Language and Cognition*, *8* (3), 369–390.

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

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2012). Concepts and categorization. *Handbook of Psychology, Second Edition, 4.*

Gómez, R. L. (2002). Variability and detection of invariant structure. Psychological Science, 13(5), 431-436.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem?. *Second Language Research*, *28*(2), 191-215.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263-277.

Hendricks, A. E., Miller, K., & Jackson, C. N. (2018). Regularizing unpredictable variation: Evidence from a natural language setting. *Language Learning and Development, 14*(1), 42-60.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290-304.

Hoedemaker, R. S., Ernst, J., Meyer, A. S., & Belke, E. (2017). Language production in a shared task: Cumulative Semantic Interference from self-and other-produced context words. *Acta psychologica*, *172*, 55-63.

Hoff, E. (2015). Language Development. In M. H. Bornstein & M. E. Lamb (Eds.), *Developmental Science* (7th ed., pp. 443-488). Psychology Press.

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological science*, *29*(6), 961-971.

Hopman, E.W.M. & Zettersten, M. (2018). Immediate feedback is critical for learning from your own productions. *Psycholinguistics in Flanders*, Ghent University, Belgium.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, *1*(2), 151-195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, *59*(1), 30-66.

Jacobs, C. L., Cho, S. J., & Watson, D. G. (2019). Self-priming in production: evidence for a hybrid model of syntactic priming. *Cognitive Science,* 43, e12749.

Kang, S.H.K., Gollan, T.H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin Review, 20,* 1259-1265

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Kapnoula, E. C., & Samuel, A. G. (2022). Reconciling the contradictory effects of production on word learning: Production may help at first, but it hurts later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(3), 394.

Karpicke, J. D. & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science, 331,* 772-775

Keppenne, V., Hopman, E. W. M., & Jackson, C. N. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics*, *42*(4), 907-936.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology, 54* (4), 283-318.

Krashen, S. (2003). *Explorations in Language Acquisition and Use*. Portsmouth: Heinemann.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589-600.

Kuczaj, S. A. (1978). Children's judgments of grammatical and ungrammatical irregular past-tense verbs. *Child Development*, 319-326.

Levelt, W. J. M. (1993). *Speaking: From intention to articulation*. MIT press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585-589.

Macdonald, D., Yule, G., & Powers, M. (1994). Attempts to improve English L2 pronunciation: The variable effects of different types of instruction. *Language Learning*, *44*(1), 75-100.

MacDonald, M. C. (2013a). How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, 226.

MacDonald, M. C. (2013b). Production is at the left edge of the PDC but still central: response to commentaries. *Frontiers in psychology*, *4*, 227.

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, *25*(1), 47-53.

MacIntyre, P. D. (1999). Language anxiety: A review of the research for language teachers. In D. J. Young (Ed.), *Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere*, (pp. 24-41). Boston: McGraw Hill

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*, 390-395.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, i-178.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language learning, 53*(1), 167-210.

Masur, E. F. (1995). Infants' early verbal imitation and their later lexical development. *Merrill-Palmer Quarterly (1982-),* 286-306.

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-Starting Item Parameters for Adaptive Language Tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 883-899).

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516-522.

Meade, G. (2020). The role of phonology during visual word learning in adults: An integrative review. *Psychonomic Bulletin & Review, 27* (1), 15-23.

Mirković, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PloS one, 11*(4), e0152489.

Oded, B., & Walters, J. (2001). Deeper processing for better EFL reading comprehension. *System*, *29*(3), 357-370.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *In Proceedings of the annual meeting of the cognitive science society* (Vol. 26, No. 26).

Pallier, C., Dehaene, S., Poline, J. B., LeBihan, D., Argenti, A. M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity in adopted adults: Can a second language replace the first? *Cerebral Cortex, 13*, 155–161.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin, 144*(7), 710-756.

Paul, J. Z., & Grüter, T. (2016). Blocking effects in the learning of Chinese classifiers. *Language Learning, 66*(4), 972-999.

Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language, 84*, 108–127.

Perek, F., & Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition, 168*, 276–293.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43-102.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760-793.

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

Ribot, K. M., Hoff, E., & Burridge, A. (2018). Language use contributes to expressive language growth: Evidence from bilingual children. Child development, 89(3), 929-940.

Robenalt, C., & Goldberg, A. E. (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, *26*(3), 467-503.

Robenalt, C., & Goldberg, A. E. (2016). Nonnative speakers do not take competing alternative expressions into account the way native speakers do. *Language Learning*, *66*(1), 60-93.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, *17*, 249-255.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. Cognitive science, 37(7), 1290-1320.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3), 1.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Vol 2: Psychological and Biological Models.* Cambridge, MA: MIT press.

Sanjeevan, T., Rosenbaum, D.A., Miller, C., van Hell, J.G., Weiss, D.J., & Mainela-Arnold, E. (2015). Motor issues in specific language impairment: a window into the underlying impairment. *Current Developmental Disorders Rep*, *2*, 228-236.

Sanli, E. A., Patterson, J. T., Bray, S. R., & Lee, T. D. (2013). Understanding self-controlled motor learning protocols through the self-determination theory. *Frontiers in psychology*, *3*, 611.

Scheffé, H. (1999). *The analysis of variance* (Vol. 72). John Wiley & Sons.

Schramm, P., & Rouder, J. (2019, March 5). Are reaction time transformations really beneficial? https://doi.org/10.31234/osf.io/9ksa6

Schwab, J. F., Lew-Williams, C. & Goldberg, A. E. (2018). When regularization gets it wrong: children over-simplify language input only in production. *Journal of child language*, *45*(5), 1054-1072.

Schwering, S. C. & MacDonald, M. C. (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in human neuroscience,* 14, 68.

Segaert, K., Kempen, G., Petersson, K. M., & Hagoort, P. (2013). Syntactic priming and the lexical boost effect during sentence production and sentence comprehension: An fMRI study. *Brain and language*, *124*(2), 174-183.

Seidenberg, M. S., & MacDonald, M. C. (2018). The impact of language experience on language and reading. *Topics in Language Disorders, 38* (1), 66-83.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus Production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning, 63* (2), 296-329.

Sim, Z. L., Tanner, M., Alpert, N. Y., & Xu (2015). Children learn better when they select their own data. In P. P. Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio (Ed.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2194–2199). Austin, TX: Cognitive Science Society.

Stroh, E. N. (2012). *The effect of repeated reading aloud on the speaking fluency of Russian language learners.* Unpublished doctoral dissertation, Center for Language Studies, Brigham Young University, Utah.

Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of research in second language teaching and learning* (pp. 495-508). Routledge.

Tachihara, K., & Goldberg, A. E. (2019). Reduced competition effects and noisier representations in a second language. *Language Learning.*

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: the lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*(1), 135-140.

VanPatten, B. (2004). *Processing instruction: Theory, research, and commentary*. Routledge.

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology*, *108*(1), 1-27.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological science*, *25*(7), 1314-1324.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776-806.

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*(3), 458-478.

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.

Woollams, A. M., Joanisse, M., & Patterson, K. (2009). Past-tense generation from form versus meaning: Behavioural data and simulation evidence. *Journal of Memory and Language*, *61*(1), 55-76.

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, *154*, 102-117.

Young, D. (1990). An Investigation of Students' Perspective on Anxiety and Speaking. *Foreign Language Annals, 23,* 539-553

Zettersten, M. (2018). Active learning: information-seeking strategies across development. (Preliminary Examination Paper)

Zaragoza, M. S., Belli, R. F., & Payment, K. E. (2006). Misinformation effects and the suggestibility of eyewitness memory. *Do justice and let the sky fall: Elizabeth F. Loftus and her contributions to science, law, and academic freedom*, 35-63

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, *124*, 342-349.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340-352.

Appendix A

Full set of Language and Visual Stimuli

Table A1

All Artificial Language Words

Determiners	**Large neighborhood (18)**		**Small neighborhood (12)**	
	singular	*plural*	*singular*	*plural*
dap	roozok	roozool	pexesh	pexaaf
jeb	monok	monool	glarmesh	glarmaaf
ked	kredok	kredool	stamesh	stamaaf
lom	frusok	frusool	gupesh	gupaaf
	hullok	hullool	zimbesh	zimbaaf
	biffok	biffool	zoutesh	zoutaaf
	pavok	pavool	pumesh	pumaaf
	teepok	teepool	mipesh	mipaaf
	wimok	wimool	kinesh	kinaaf
	foudok	foudool	heefesh	heefaaf
	chagok	chagool	buresh	buraaf
	chufok	chufool	chetesh	chetaaf
	ditok	ditool		
	sarbok	sarbool		
	plimok	plimool		
	naosok	naosool		
	zeevok	zeevool		
	skunok	skunool		

Note. Stems are listed with suffixes for singular on the left and for plural on the right, organized by neighborhood.

Figure A1
Overview of all Visual Stimuli

Note. The 18 large neighborhood aliens (depicted on the left) are humanoid: they stand upright on two legs and have two arms, giving them a tall appearance. The 12 small neighborhood aliens (depicted on the right) are blob-like – while they have faces, they do not have legs or arms and thus have a more wide, squat shape. This overview picture of all aliens was shown to participants at the start of the study while they listened to a 1:05 minute audio introduction to the purpose of the study.

Appendix B

Data Filtering

Table B1 contains a full overview of all the data I gathered, and which data I did and didn't include in our analyses as reported in the main text. The full process for filtering data is described in the text of this appendix. Note that the process of data filtering was somewhat iterative, and that a participant I excluded for one reason would often also have been excluded for other reasons. Thus, Table B1 only lists the reason I marked when excluding the participant.

Table B1

Breakdown of Participant Numbers by Condition and Data Categorization.

	category	reason	p	c	NA	total
unusable	personal	unfocused browser	20	19	-	39
		ended early	13	15	4	32
		did not try	14	11	-	25
		distracted/interrupted	2	7	5	14
		other (took notes, sick, etc)	6	6	1	13
		total personal	55	58	10	123
	unknown (did not finish)	no server access	-	-	17	17
		server accessed	32	11	41	84
		total unknown	32	11	58	101
	technical	experiment crashed	10	5	6	21
		device / screen size	1	-	20	21
		safari	2	3	7	12
		some content didn't show	21	26	1	48
		total technical	34	34	34	102
	performance	SC threshold	5	1	-	6
		2+ thresholds	14	15	-	29
		SDT threshold	0	2	-	2
		TC threshold	2	2	-	4
		FC threshold	7	3	-	10
		total performance	28	23	-	51
	total unusable		149	126	102	377
partially usable		personal	6	3	-	9
		technical	2	4	-	6
		performance: EM-threshold	7	14	-	21
	total partially usable		15	21	-	36
fully usable			103	125	-	228
total provided consent			267	272	102	641

Preregistered Data Filtering

Unusable and Partially Usable Data

If a participant did not access any tests, they were automatically categorized as unusable, since there was no test data to analyze. If participants started the tests but did not finish at least both the two alternative and four alternative forced choice tests, they were also immediately categorized as unusable. This is because I had planned to use data from the four alternative forced choice test as a covariate in further analyses.

If a participant finished both forced choiced tests, but not the error monitoring and production tests, I considered their data for the two forced choice tests usable. If participants

only finished part of the error monitoring test, I did not consider the error monitoring test data as usable. If participants finished the error monitoring test but not the production test, I considered their error monitoring data usable, as well as whichever trials of the production test they completed. Thus, the only test on which I considered partial data as analyzable was the production test. The reason for this is that it was the final test, and so their learning during this test (or missing test items) wouldn't affect further tests.

Sometimes participants did finish a test (or the whole experiment), but missed certain trials during a test. For example, some participants indicated in the post-experiment survey (see Appendix C) that during certain trials, an image hadn't loaded on the server. If a participant missed content either during training or either of the two forced choice tests, their data was automatically considered unusable. If they missed content during the error monitoring test, their data up until that point was considered usable but any later trials were not analyzed (so this could lead to analyzing only the first half of a participant's error monitoring test data). If they missed content during the production test, only those trials they missed were not analyzed.

Classifying Reasons for Unusable and Partially Usable Data

Many SONA participants provided consent on Qualtrics and filled out the language background survey, but did not finish (or even start) the experiment on the lab server. Whenever a participant provided consent but didn't finish the experiment, I checked the data on the server to see if I could deduce a reason for this (e.g. server crash). I also checked to see whether the participant had failed to resume the experiment within 5 minutes during scheduled breaks. I also checked browser information, in order to see whether the experiment was conducted in a browser with known issues playing the experiment (Safari) or on a screen that was too small to properly display the experiment (e.g. a smartphone; minimum resolution was 1275 x 680). Finally, if available, I consulted the post-experiment qualtrics survey (see Appendix C) to check if participants had mentioned problems there, or if they had indicated not trying their best, in which case I also excluded their data.

If I was not able to learn what happened, I followed up with a brief email to ask how much time the participant spent (so that I could give them fair credit for the time they spent on the experiment) and to ask if there had been any technical difficulty with the experiment.

Finally, the server also noted when participants unfocused their browser during the experiment. If this happened for longer than 10 seconds during experimental content, the data were categorized as unusable from that point onward (following the same prior rules about e.g. needing to have usable data for both forced choice tests for data to be usable).

Additional, Post-hoc Filtering

After filtering and categorizing the data from online participants in these preregistered ways, the reaction time data in particular were still really poorly distributed, with a lot of noise and reaction times that were very long or very short (e.g. ~100 ms). I also saw people who e.g. always pressed one key during the error monitoring test, regardless of the trial. I concluded that the online data I had gathered was still too noisy to analyze, and decided to threshold the data in several additional ways.

Accuracy

The experiment included 18 sound check trials, in which participants heard an English word and had to type that word into a box. I automatically marked a participant's data as unusable if they got fewer than 83% (15/18) of these sound check trials correct. If a participant had 3 errors, I checked the errors by hand. If 2-3 of these errors were real errors (and not typos),

the participant's data was marked as unusable. This is listed as SC (Sound Check) threshold in Table B1.

In the signal detection analysis of the error monitoring data, besides generating a d' score, I also generate a bias score *c*. This bias score indicates whether a participant was more likely to press one specific button. Based on the distribution of bias scores, 5 participants looked like outliers, with $c > 0.5$ or $c < -1.5$. I manually looked at the the data of the participants inside the boundaries closest to these outliers to see if there was evidence that these non-outliers also mainly used one button repeatedly, but this was not the case. This is listed as SDT (Signal Detection Theory) threshold in Table B1.

Reaction Time

I re-processed reaction time data for each comprehension test (Two alternative forced Choice: TC; Four alternative forced Choice FC, Error Monitoring EM). This time, I first put in absolute boundaries of 0.3-8s, and marked trials that fell outside of that boundary. Then, I calculated for each test a participant's mean and standard deviation on trials that were left inside these boundaries and answered correctly. I marked any trials (both correctly and incorrectly answered) with reaction times outside of that boundary. Then, I calculated for each participant what percentage of their trials for each test was marked as being outside of boundaries. Based on distributions, I noted that most participants contributed at least 75% of trials per test that were within these boundaries. Thus, I decided to mark a participant as below threshold on a test if <75% of reaction times for that test was outside of these boundaries.

Performance Thresholding

To summarize, I had 5 post-hoc performance thresholds: SC & SDT (accuracy based) and TC, FC and EM (reaction time based). I first excluded participants who did not meet the sound check threshold. Then I excluded participants who failed two or more thresholds. I then excluded participants who failed the SDT, TC and FC thresholds. Finally, for participants who only failed the EM threshold, I marked their TC and FC data as usable, but their EM and production test (because it was afer the EM test) data as unusable, thus making their data partially usable. The majority of participants who were in the dataset after excluding participants only for preregistered reasons passed all of these thresholds and were categorized as fully usable.

Note that this reaction time thresholding was separate from the main reaction time analyses. Thus, in the main reaction time analyses, I did e.g. include a trial with a 12 s reaction time if it is within that participant's mean + 3SD for correct trials for that test. The reaction time thresholds mentioned in this appendix were only used to assess, overall, whether a participant provided enough data with 'reasonable' reaction times or whether they provided so much data outside of reasonable reaction time boundaries that I had good reason to believe they may have not fully been paying attention to the experiment.

Partially Usable Data

Table B2 lists, for participants with partially usable data for any of the reasons mentioned, how much analyzable data for each test I have.

Table B2

Breakdown of Participant Numbers by Condition and Which Test Data Were Usable

what data is usable?	p	c	total
only TC & FC	12	18	30
FC, TC & some EM	1	1	2
FC, TC, EM & some PT	2	2	4
total	15	21	36

Condition Assignments

Participants were randomly assigned to a condition through the pre-experiment qualtrics script, with both conditions equally likely initially. Note that, if a participant did not finish at least the first block of training (8 passive trials and 8 active trials), they are listed in Table B1 as condition 'NA'. Many of these participants never accessed the server, and those who did access the server may never have seen the 8 active trials at the end of the first block of training. Thus, even though the qualtrics survey assigned a condition to these people, I counted them as not having had a condition assignment.

I preregistered that, if I noticed unequal attrition between conditions, I would change condition assignments. Based on piloting, I expected to see more attrition in the production condition than in the comprehension condition. On November 20th, 2021, when 342 participants had signed up, provided consent and their participation deadline on SONA had passed, I assessed by-condition attrition and changed condition assignment to be 3:1 production:comprehension based on the data in the bottom two rows in Table B3.

Note that even after that date, participants still had some likelihood of being assigned to the comprehension condition, so that I kept sampling both conditions over time. Finally, note that I still ended up with more comprehension data (125 fully usable, 21 partially usable) than production data (103 fully usable, 15 partially usable). Thus, in order to have ended up with balanced data numbers between conditions, I could have set the likelihood differents of assignment more in favor of production, or changed the likelihoods earlier in the running process.

Table B3

Breakdown of Participant Numbers on November 20th, 2021 by Condition and Initial/Final Usability

initial categorization	final categorization	p	c	NA	total
usable	fully usable	63	81	-	144
	partially usable	8	5	-	13
	unusable	17	27	-	44
	total initially usable	88	113	-	201
	unusable	54	30	57	141
total numbers of participants run by November 20th		142	143	57	342

Note. Initial categorization was done based on data I had preregistered to access before datacollection was done. So, at this point I had only accessed the following reasons a participant's data might be classified as unusable: unknown (ended early, with or without server access, without reason given); some technical reasons (experiment crashed, safari, screen size/device); some personal reasons (ended early, or, within 'other', if a participant had e.g. emailed me to let me know they were sick). At that time, the other reasons for classifying data as unusable were not accessed yet: all performance exclusion criteria, some technical reasons (content missed), and some personal reasons (did not try, unfocused browser, or noted in post-experiment survey that they were interrupted or distracted, or reasons like taking notes that fell under 'other').

Appendix C

Pre- and Post-Experiment Surveys

For all multiple choice questions, answers were scrambled into a random order and no responses were required, so participants could skip answering questions if they wanted to. Multiple choice answer options are listed alphabetically here inside {curly brackets}, options are separated by a semi-colon. When a text box was provided as a general response option this is indicated as {text box}, or as a clarification option when certain multiple choice options were selected this is indicated as {{text box}}. Information inside [square brackets] pertains to survey flow and display logic.

Pre-Experiment Survey

Meta-data & informed consent:

[At the start of the survey, meta-data about browser, operating system and screen resolution was automatically detected. If Safari was detected, a message was displayed that directed the participant to restart the survey in a different browser if they wanted to participate in the experiment. If safari was not detected, my contact information was displayed (in order to be able to contact me in case of technical difficulties with the experiment), and on the next page the informed consent screen was displayed. Once a participant provided informed consent, the following questions were displayed on a new page.]

Demographic questions:

1. How old are you.
   {text box}
2. What is your gender?
   {Man; Other (please list) {{text box}}; Woman; Would rather not say}
   [page break]
3. Which option(s) best describe your race and/or ethnicity (check all that apply)?
   {American Indian / Native American; Asian; Black / African American; Hispanic / Latino; Other (please list) {{text box}}; Pacific Islander; White / Caucasian; Would rather not say}
   [page break]

Language background questions:

4.
   a. Are you a native English speaker?
      {no; yes}
   b. [only displayed if 'no' was selected on 4.a.]
      If not English, what is your native language?
      {text box}
      [page break]
5.
   a. Did you learn other languages than your native language at home before age 5?
      {no; yes}
   b. [only displayed if 'no' was selected on 5.a.]
      What other language did you learn at home before age 5? Are you bilingual?
      {text box}
      [page break]

6.
    a. Did you learn any languages later in life? (ex: at school or an after-school program)
    {no; yes}
    [if no was selected, participant progressed immediately to 8.]
        b. What is the language you know best besides your native language(s)? (you will get asked about other languages you might know later on)
        {text box}
        c. How many years of experience do you have with that language?
        {1-3 years; 3-5 years; 5-up years}
        [page break]
7. [this question was displayed up to five times in a loop]
    a. Apart from the languages you already mentioned, did you learn any other languages later in life? (ex: at school or an after-school program)
    {no; yes}
    [if no was selected, participant progressed immediately to 8.]
        b. What is your [third, fourth, …] language?
        {text box}
        c. How many years of experience do you have with that language?
        {1-3 years; 3-5 years; 5-up years}
        [page break]
8. This experiment can only be done on a desktop or laptop computer, it will not work on phone or tablet screens, since those screens are too small to display the experiment properly. Please confirm that you are using a desktop or laptop computer.
    {no, I am using a phone or tablet; yes, I am using a desktop or laptop computer}
9. This experiment includes sound, and takes concentration. While the experiment can be done remotely, we want you to treat it as you would an in-person experiment. That means we expect you to be alone, in a quiet room, while doing this experiment. Just like for an in-person experiment, your phone should be on silent and not within reach. You should not have any background noise on (tv, radio, etc). Please confirm that you are alone in a quiet room without distractions.
    {no, I am not alone in a quiet room without distractions; yes, I am alone in a quiet room without distractions}

[If ‘no’ was selected for either 8. or 9., the participant was directed to come back and complete the experiment at a later moment, when they had access to a laptop or desktop computer in a quiet room without distractions. Participants were offered my contact information in case they did not have access to this, so that they could opt to complete the experiment in our lab space if they never had access to this, or extend their SONA deadline if they had access but not within the deadline they had signed up for. Otherwise, participants were automatically redirected to the lab server where the experiment started.]

Post-Experiment Survey

This survey asked, with questions progressing from open to closed, about the regularities participants had learned, in order to survey explicit awareness. These data were gathered for the benefit of second language acquisition researchers who are interested in questions about explicit versus implicit awareness of grammar learning. Since this was a not a research question in this study, they were not coded or analyzed. These (anonimized) responses will be made publicily available together with language learning data in case SLA researchers wish to further analyze them.

Open questions:

1. What did you notice about how the language you learned worked?
   {text box}
   [page break]
2. Wat did you notice about how the aliens looked?
   {text box}
   [page break]
3. What did you notice about how the language you learned was related to how the aliens looked?
   {text box}
   [page break]

Closed questions:

4.
   a. Did you notice that there was a difference between singular and plural in the language?
      {no; no, I didn't during the study, but now that you mention it, I think I know what the difference was; yes}
   b. [not displayed if 'no' was selected for 4.a.]
      What was it you noticed about singular and plural?
      {text box}
      [page break]

5.
   a. Did you notice that different groups of aliens had different shapes?
      {no; no, I didn't during the study, but now that you mention it, I think I know what the difference was; yes}
   b. [not displayed if 'no' was selected for 5.a.]
      What was it you noticed about their shapes?
      {text box}
      [page break]

6.
   a. Did you notice that the language referred differently to the two different groups of aliens with different shapes?
      {no; no, I didn't during the study, but now that you mention it, I think I know what the difference was; yes}
   b. [not displayed if 'no' was selected for 6.a.]
      What was it you noticed about how the language differently referred to the two differently shaped groups of aliens?
      {text box}

[page break]

Screening questions (to determine if data was valid):

7. Any technical difficulties during the study?
   {text box}
8. Did the audio work while you did the experiment for both the English words and the new language?
   {text box}
   [page break]

[The next page showed a debrief of the purpose of the study. After that, there was a page with confirmation that the experiment was now complete, data was sent to the server, and my contact details if they had any other questions. Below that, two more questions were displayed.]

9. Did you do your best to actually learn the language in a quiet environment without distractions? Note that you wil be automatically redirected to receive credit, and the answer you give here will not affect your credit.
   {I did my best and was in a quiet space without distractions; I did my best but there were distractions and/or noise in the space I was in; I did not do my best, and there were distractions and/or noise in the space I was in; I did not do my best but I was in a quiet space without distractions; Other, namely {{text box}}}
10. Any other information that you want to share that might affect how you did the study? E.g. were you not sober, did you cheat by taking notes, were you incredibly sleep deprived, anything at all that might affect your learning ability? Note that this information is anonymous and will not be linked to your name, SONA ID or other identifying information, and will not be shared outside the research team.
    {text box}

Appendix D

Production data processing

Automated Initial Processing

Production Attempt -> Parsed Determiner and Noun

A script first parsed each production attempt (a participant's typed response for a production trial) into two words: a determiner and a noun.

- If a participants typed two words (word = set of letters, demarcated by whitespace), the first word was interpreted as the determiner, and the second word was interpreted as the noun.
- If only a single word was typed, this word was repeated and parsed both as a determiner and a noun. This was done because when a participant typed a single word, it was sometimes a determiner, and sometimes a noun.
- If three words were typed, the second and third were pasted together to form a single word and were processed together as the noun; this was done because there were cases where participants typed a space between the stem and the suffix.
- Finally, if four or more words were typed, the script scored the utterance as unparseable and fully incorrect, and did not further process it.

Thus, unless an utterance was deemed unparseable, the output of the first parsing step was always a determiner and a noun.

Parsed Noun -> Parsed Stem and Suffix

The second step of the script parsed the second word into a stem and a suffix. It did so by splitting up the noun after the first consonant cluster (1 or more consonants) to follow the first vowel cluster (1 or more vowels). All stems in the artificial language took the form 'consonant cluster – vowel cluster – consonant cluster', and all suffixes took the form 'vowel cluster – consonant cluster'. Thus, all correct artificial language nouns were separated cleanly into the stem and the suffix this way. Note that on trials where a participant only typed a stem, this would lead to the script concluding that no suffix was present. Thus, after the second step, the noun was parsed into a stem (always present) and (where present) a suffix

Parsed Determiner, Stem and Suffix -> Assigned Artificial Language Morphemes (1)

The third step of the script attempted to assign each parsed morpheme to an artificial language morpheme of the same type (e.g. a parsed determiner could only be assigned an artificial language determiner, not an artificial language stem). The script calculated the Levenshtein Distance (LD) between the produced morpheme and every legal morpheme of that type in the artificial language (4 for determiners & suffixes, and 30 for stems). An LD threshold of 4 was set for stems. An LD threshold of 3 was set for determiners and suffixes. Legal determiners and suffixes in the artificial language had a maximum of 3 characters, so LD 3 or above would mean an attempt either had no characters in common with any legal morpheme of that type, or was at least double the length of any legal morpeheme of that type.

- If no LD was strictly below threshold, the parsed morpheme was scored as incorrect and not processed further.
- If at least one LD was below threshold:
    - If a single artificial language morpheme of that type had the lowest LD, the parsed morpheme was assigned as that artificial language morpheme (e.g. both 'kred' and 'kret' were assigned as legal artificial language stem 'kred').

- If several artificial language morphemes had equal lowest, below-threshold LD's, the parsed morpheme was marked for hand-coding and the parsed morpheme and its nearest LD neighbors were written into a separate file.

At the end of this third step, all morphemes were thus either scored as incorrect (at or above threshold), marked as hand-coding needed (equidistant below-threshold LD's) or assigned as a specific attempted morpheme in the artificial language.

Human Coding

A total of 243 parsed determiners, 2179 parsed stems and 171 parsed suffixes were marked for hand-coding because they had lowest, equidistant, below-threshold LD (equi-LD) to several artificial language morphemes of their type. Note that for efficiency reasons production data processing was happening in parallel with general data filtering (Appendix B). Thus, the numbers in this section are based on both data categorized as usable and some data (later) categorized as unusable. Furthermore, even though I analyze only production test data, productions from active production training trials were also processed and are included in these numbers. Processed active production training data will be made available on OSF for interested researchers to further analyze.

The coders were 4 native English speakers, all of whom were familiarized before coding with the experiment reported here and the artificial language used. During coding, coders had access to Table A1 with all morphemes in the artificial language used in the experiment written out, as well as the sound recordings of the artificial language used in the experiment. Coders were simply presented with lists of parsed morphemes, and for each parsed morpheme the equi-LD artificial language morphemes output by the script. Thus, coders did not have access to participant, condition or trial information. Lists were labeled by morpheme type, so coders did know whether a morpheme was parsed by the script as a determiner, stem or suffix. The stems list was split up into smaller lists. Lists were sent out to the coders until every list had been coded by 3 coders.

Coders were instructed to code each parsed morpheme as either 0 or 1, with 0 indicating no further processing was possible (either because the parsed morpheme was a keyslam, an English word, not like any of the morphemes in the artificial language, or too ambiguously in between two morphemes of the artificial language) [1]. If a coder indicated 1, they also indicated which artificial language morpheme (or which concatenation of 2 morphemes) they thought the parsed morpheme should be assigned to. In most cases, this was an artificial language morpheme from the same type as the parsed morpheme. Coders were allowed to assign artificial language morphemes as the intended target that weren't equi-LD or weren't from the same type as the parsed morpheme and thus weren't listed by the script. This was relevant in several types of cases. Sometimes, participants used a determiner as a stem (e.g. 'dapok', would lead to the parsed stem 'dap' which has several equi-LD artificial language stems but is identical to an artificial language determiner). Sometimes, participants had missed a consonant due to a typo (e.g. 'chaool', which would be parsed as a stem without a suffix, but is also a typo of 'chagool').

Coder responses were then compared. If no coders or only a single coder coded a parsed morpheme as 1, it was assigned 'none'. If 2 or 3 coders coded a parsed morpheme as 1, it was processed further. If 2 or 3 coders provided the same assigned artificial language morpheme (or

[1] Note that initially, coders were given more different categories for the different instances mentioned here; however, once I realized all of these would not be processed further and simply marked as incorrect, I simplified the categories to make hand-coding easier and faster.

concatenation of 2 morphemes), the parsed morpheme was assigned as that artificial language morpheme. If all 3 coders listed a different assigned artificial langauge morpheme, the parsed morpheme was mapped to 'none'. Finally, if only 2 coders had coded '1', but they provided different assigned artificial language morphemes, the fourth coder (or third[2]) was invoked as a tiebreaker, and definitively assigned the morpheme either to one of the two provided ones or to 'none'. Assignment outcomes, split out by coder agreement is listed in Table D1 for parsed determiners, in Table D2 for parsed stems and in Table D3 for parsed suffixes. Joint together, all of the parsed morphemes were added into a 'dictionary' listing the parsed morpheme, morpheme type (determiner, stem, suffix), and assigned artificial language morpheme(s) (or, 'none').

Table D1

Hand-Coding Outcomes for Parsed Determiners

assignment	full agr.	majority agr.	tiebreaker	full disagr.	total
'none'	130	82	2	0	214
A.L. morpheme	3	21	5	-	29
total	133	103	7	0	243

Note. A.L.: Artificial Language; agr.: agreement; disagr.: disagreement.

Table D2

Hand-Coding Outcomes for Parsed Stems

assignment	full agr.	majority agr.	tiebreaker	full disagr.	total
'none'	1082	733	18	6	1839
A.L. morpheme	47	218	75	-	340
total	1129	951	93	6	2179

Note. A.L.: Artificial Language; agr.: agreement; disagr.: disagreement.

Table D3

Hand-Coding Outcomes for Parsed Suffixes

assignment	full agr.	majority agr.	tiebreaker	full disagr.	total
'none'	71	68	2	1	142
A.L. morpheme	5	19	5	-	29
total	76	87	7	1	171

Note. A.L.: Artificial Language; agr.: agreement; disagr.: disagreement.

Automated Final Processing

Parsed, Equi-LD Morphemes -> Assigned Artificial Language Morphemes (2)

The dictionary mapping parsed morphemes onto either 'none' or an assigned artificial language morpheme was then integrated into the initial script. Note that, whenever the dictionary assigned two morphemes for one parsed morpheme, the two assigned morphemes were always

[2] Preference was given to the fourth coder, who had never seen this attempted morpheme yet. However, not all coders were available anymore at this stage of coding, so if the fourth coder was not available, the third coder (who had initially assigned 0) was assigned as a tiebreaker instead.

stem + suffix. The trials where the dictionary assigned two artificial language morphemes for one parsed morpheme boiled down to three categories[3].

- The parsed morpheme was a stem, and there was no parsed suffix, in which case the two assigned morphemes were simply assigned to the stem and suffix. For example, the attempted production 'jeb bresh' was parsed by the script into 'jeb' (determiner), 'bresh' (stem), '' (suffix). The parsed stem 'bresh' was then assigned 'bur' (stem), 'esh' (suffix) in the dictionary by the coders. On the second run, the script is able to parse and assign this attempted production as 'jeb' (determiner), 'bur' (stem), 'esh' (suffix).
- The parsed morpheme was a stem, and there was also a parsed suffix, in which case the stem was assigned to 'none'. For example, the attempted production 'jeb breshok' was parsed by the script into 'jeb' (determiner), 'bresh' (stem), 'ok' (suffix). The parsed stem 'bresh' was then assigned 'bur' (stem), 'esh' (suffix) in the dictionary by the coders. On the second run, the script parses and assigns this attempted production as 'jeb' (determiner), 'none' (stem), 'ok' (suffix).
- The parsed morpheme was a determiner or suffix, in which case that morpheme was assigned to 'none'. For example, the attempted production 'bresh chagok' was parsed by the script into 'bresh' (determiner), 'chag' (stem), 'ok' (suffix). The parsed determiner 'bresh' was then assigned 'bur' (stem), 'esh' (suffix) in the dictionary by the coders. On the second run, the script parses and assigns this attempted production as 'none' (determiner), 'chag' (stem), 'ok' (suffix).

The script was re-run, and now each parsed morpheme was either:

- scored as 'incorrect', if at or above LD threshold, no response, 'none', or morpheme of the wrong type (e.g. 'dap' as a stem)
- or assigned to an artificial language morpheme of the appropriate category.

Assigned Artificial Language Morphemes -> Score (Correct/Incorrect)

If an artificial language morpheme was assigned, it was checked whether this was the correct target morpheme to describe the picture shown in that production trial (scored as 'correct') or not (scored as 'incorrect').

Assigned Artificial Language Morphemes -> Grammatical Category

If an artificial language morpheme was assigned, the script retrieved grammatical category information (neighborhood for all morphemes, and plurality for determiners and suffixes). This way, I could analyze whether an identifiable but incorrect morpheme crossed category boundaries (potential overgeneralizations).

---

[3] The third category was easy to code and implemented in the script. Since the first two categories only consisted of a total of 35 trials, I chose to assign these by hand instead of implementing the described procedure in the script.

Appendix E

Predictions from dissertation proposal

This section is copied verbatim from the dissertation proposal I submitted to my committee in November 2020, with the exception of figure labels. Also note that I refer to familiarity (familiar/unfamiliar) as 'alien type' (trained/untrained), and to the large neighborhood as the 'big' neighborhood in this section copied from my dissertation proposal.

In Hopman & MacDonald (2018), we hypothesize that production training leads to better learning of grammatical features than comprehension learning. Specifically, during production planning, different features relevant to the to-be-produced utterance are held together in working memory, allowing for binding between these features. In that experiment, we only tested grammar comprehension of novel combinations of well-trained lexical items. We found that production participants performed better on those than comprehension participants, which is in line with stronger binding between the grammatical, lexical and visual features (e.g. the specific visual features of the scary-looking alien named 'teep', the visual features of scary-looking aliens in general, the 'teep' stem for that alien, and the '-us(u)' suffixes for scary-looking aliens). Note that, while I talk about all of these as 'binding' to emphasize that they are all happening at the same time, they are typically thought of as different types of learning. For example, learning that a specific scary-looking alien is part of a category of aliens that share certain features is category learning (see e.g. Goldstone et al., 2012 for a review), while learning that a particular lexical stem is often followed by one of two suffixes is sequential learning (see e.g. Seidenberg & MacDonald, 2018 for a review), and mapping a unique visual stimulus to a lexical stem is word learning (see e.g. Meade, 2020 for a review). What is unclear from the Hopman & MacDonald experiment is how lexically specific this binding of different types of features is, and whether or not it would extend to production training improving the binding between a category of grammatical features and a set of visual features, *in absence of the specific (lexical) item*. This empirical question will be answered by the current experiment, but before I dive into two diverging sets of predictions for how production training might affect generalization (an interaction effect), I will first establish overall predictions independent of any interactions with learning condition.

Based on prior literature, I have the following predictions for the three main independent variables (learning condition, neighborhood size, alien type):

- For the main effect of learning condition, I expect to find that production-trained participants score higher overall than comprehension-trained participants (Hopman & MacDonald, 2018).
- For the main effect of neighborhood size, I expect to find that participants score higher overall on items testing big neighborhood aliens than items testing small neighborhood aliens. While each indivual alien has been seen equally often in training, since there are more big neighborhood aliens, that type frequency and thus practice with its grammatical elements (determiners and suffixes) is higher, which should make it easier to get those items correct on a test (Keuleers et al., 2007).
- For the main effect of alien type, I expect to find that participants score higher overall on trained aliens than on the much less familiar test-only aliens, simply because participants have had less practice with those.
- For the interaction between neighborhood size and alien type, I expect to find that participants have an especially hard time with small neighborhood test-only

aliens, because for those aliens, both the category (small neighborhood) and the individual items (test-only) are less practiced (Keuleers et al., 2007).

These four predictions hold under each of the two different scenarios worked out below (so they are true in both figures with predictions).

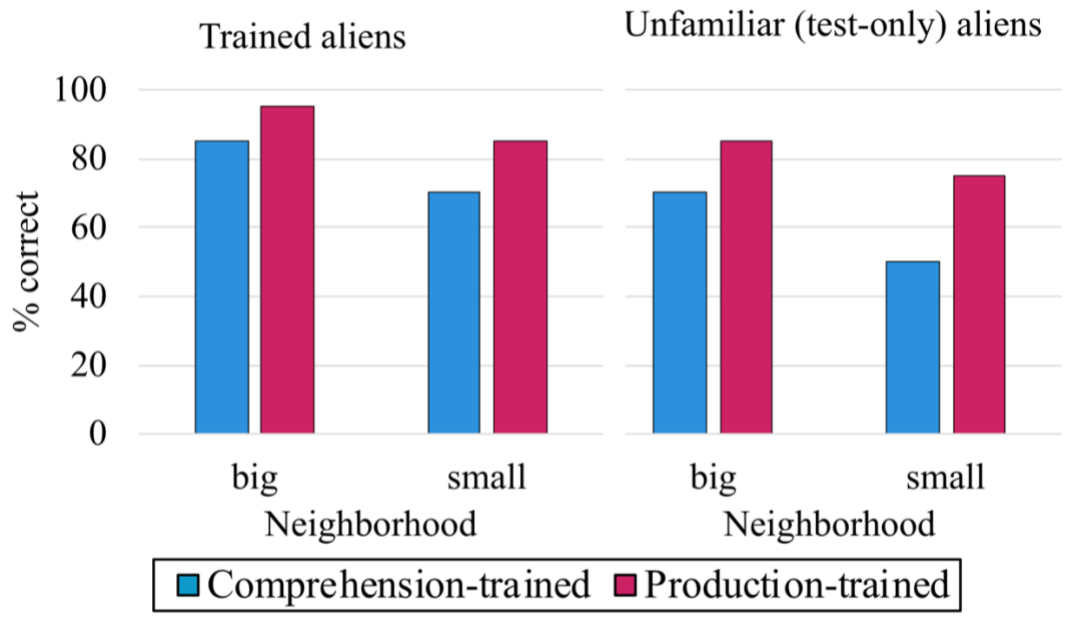
Why Could Production Training Lead to* Better *Generalization?

I hypothesize that production training leads to stronger binding than comprehension training between all four elements mentioned earlier (e.g. the specific visual features of the scary-looking alien named 'teep', the visual features of scary-looking aliens in general, the 'teep' stem for that alien, and the '-us(u)' suffixes for scary-looking aliens). This leads to the following predictions, all of which are graphed together in Figure E1.

- Specifically, in this hypothesis in the production condition there would be stronger binding between on the one hand the visual features of the big neighborhood and the suffixes and determiners used to describe them and on the other hand the visual features of the small neighborhood and the suffixes and determiners used to describe them. This leads to the specific prediction that for the two-way interaction between learning condition and alien type, production participants outperform comprehension participants on less familiar lexical items (test-only aliens).
- That same strengthened binding for production participants would also improve performance on small neighborhood aliens. Since that category of aliens is smaller, with a lower type frequency, stronger binding overall in the production condition will be especially helpful for the small neighborhood aliens. Thus, I predict that for the two-way interaction between learning condition and neighborhood size, production participants outperform comprehension participants on small neighborhood aliens.
- Finally, combining the prior two arguments, the stronger binding of category and grammatical features in the production condition will be more impactful for the less-well-learned small neighborhood, leading to the prediction that for the three-way interaction between learning condition, neighborhood size and alien type, production participants will especially outperform comprehension participants on test-only, small neighborhood aliens.

Figure E1

Predicted Outcomes Under My Hypothesis That Production Training Leads to Better Generalization



Why Could Production Training Lead to* Worse *Generalization?

However, other predictions are possible. It may be that while production practice helps learners on familiar lexical items, it hinders generalization. By the time production participants get to the tests, they are much more practiced at producing big neighborhood grammatical features (e.g. '-ok', '-ool', and the two big neighborhood determiners) than at producing small neighborhood grammatical features. This would place participants in the initial part of the U-shaped curve for learning lower frequency or irregular forms, where a more regular forms is more accessible (e.g. Ramscar & Yarlett, 2007). This asymmetry, where participants are more practiced at producing big neighborhood features, might make those grammatical features far more easily accessible and recognizable, and thus might prompt production participants to generally prefer any response option with big neighborhood features (whether e.g. in their own productions or in those features 'sounding more correct' in error monitoring trials). While participants in both learning conditions have also *heard* the big neighborhood grammatical features more often, just hearing these features in asymmetrical frequencies should create less of an over-practiced self-priming type bias than regularly *producing* those same features (Jacobs et al., 2019; Segaert et al., 2013). Thus, while all participants may make overgeneralization errors in favor of the big neighborhood (trial types where this is particularly possible were flagged in the methods section), in this scenario, production participants may be more likely to make this type of error than comprehension participants. This leads to the following predictions, all of which are illustrated in Figure E2.

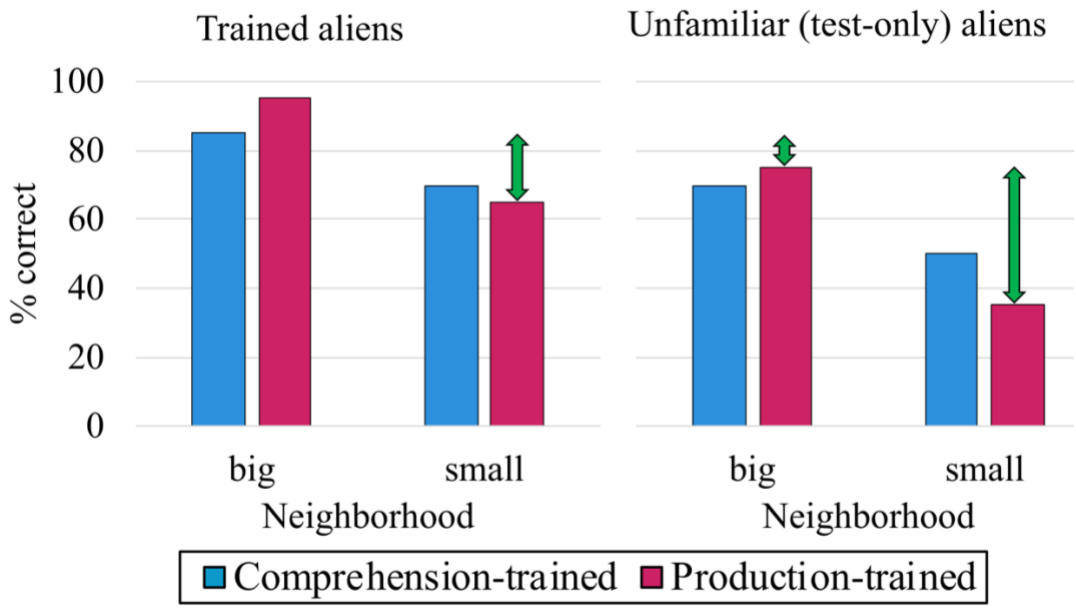
- If production-training leads participants to generally prefer answers containing big-neighborhood features, this would lead to more errors on small neighborhood features. Thus, in this scenario one would expect to find a two-way interaction between learning condition and neighbood size with production participants scoring worse than comprehension participants on small neighborhood aliens.
- In this scenario, production participants might score better than comprehension participants on (over)trained aliens, but they would do relatively worse on

unfamiliar aliens, leading to a two-way interaction between learning condition and alien type with production participants scoring relatively lower on unfamiliar test-only aliens than they did for trained aliens.

- Finally, in this scenario, one may predict a three-way interaction between learning condition, neighborhood size and alien type so that production participants do especially poorly on small neighborhoood test-only aliens. Note that in this scenario, I have graphed the prediction that production participants score below chance on small, test-only aliens. This behavior, of being systematically more likely to overgeneralize than get the correct form, is not unheard of in experiments with real language learners (e.g.Table 1; Kuczaj, 1977).

Figure E2

Predictions Under the Alternative Hypothesis That Production Training Impairs Generalization



Note. Green double arrows indicate differences with predictions under my main hypothesis (see Figure E1).

Test Modality

The two prediction graphs presented here are not for any specific test, these are my general predictions. The three different tests (forced choice, error monitoring and production) are interesting for different reasons. One possible outcome could be that the hardest test (production) shows chance or uninterpretable performance on harder aspects of the language and interesting results on easier aspects on the language that may show ceiling effects in the easier comprehension tests. Note that this is in line with the prediction I have based on my literature review that generally, the production test should show more (over)generalization errors than the two comprehension tests. I am hesitant to make more fine-grained predictions for the difference in generalization accuracy between the production- and comprehension tests. If I do find qualitative differences in the production test generalization results, it might be challenging to tease apart whether those different results are due to simply the increased difficulty of the production test itself, or whether they are due to the production modality of the test. In fact, this

might be why it was fairly hard to interpret test modality results from prior language learning studies that included both comprehension and production tests. The one more specific prediction I'm comfortable making with respect to test type is an interaction between training type and test type, so that each group should be (relatively) better on the test type that corresponds to their training type. Note that this does not mean I expect comprehension participants to do better than production participants on the comprehension tests; I just expect comprehension participants to be outperformed more by the production participants on the production test than the comprehension tests. Finally, the forced choice test, which I expect to be the easiest test, is important to include in order to establish that participants in both conditions did learn both the trained lexical items and the grammatical regularities above chance.

Appendix F

As predicted preregistration

This preregistration was uploaded on OSF.io on November 29, 2021 using the as-predicted format. I have copied it here verbatim, with the following exceptions. I have added several headings to make the structure of the preregistration clearer. I have added (sic) where I accidentally wrote 'three' instead of 'two' dependent variables for comprehension tests. I have formatted example productions and how the algorithm would parse them in a table. Also note that I refer to familiarity (familiar/unfamiliar) as 'subneighborhood' (trained/untrained), and to the large neighborhood as the 'big' neighborhood in this preregistration.

Data collection

Have any data been collected for this study already? Note: 'Yes' is a discouraged answer for this preregistration form.

It's complicated. We have already collected some data but explain in Question 8 ['Other' heading in this appendix] why readers may consider this a valid pre-registration nevertheless.

Hypothesis

Does the benefit we found for production training compared to comprehension training in earlier studies for learning grammatical features of a new language extend to novel (untrained) words? We hypothesize that production practice (with an artificial language introducing (trained) words and a novel grammatical regularity not found in English) leads to more accurate generalization to novel (untrained, introduced at test) words of the newly learned grammatical regularities than comprehension practice does. Below are the more specific predictions/hypotheses, copied from EH's dissertation proposal:

- Specifically, in this hypothesis in the production condition there would be stronger binding between on the one hand the visual features of the big neighborhood and the suffixes and determiners used to describe them and on the other hand the visual features of the small neighborhood and the suffixes and determiners used to describe them. This leads to the specific prediction that for the two-way interaction between learning condition and sub neighborhood (trained vs untrained), production participants outperform comprehension participants on less familiar lexical items (untrained aliens).
- That same strengthened binding for production participants would also improve performance on small neighborhood aliens. Since that category of aliens is smaller, with a lower type frequency, stronger binding overall in the production condition will be especially helpful for the small neighborhood aliens. Thus, I predict that for the two-way interaction between learning condition and neighborhood size, production participants outperform comprehension participants on small neighborhood aliens.
- Finally, combining the prior two arguments, the stronger binding of category and grammatical features in the production condition will be more impactful for the less-well-learned small neighborhood, leading to the prediction that for the three-way interaction between learning condition, neighborhood size and subneighborhood (trained/untrained), production participants will especially outperform comprehension participants on untrained, small neighborhood aliens.

Dependent variable

There will be four tests after training. The three comprehension tests (2AFC, 4AFC, EM) all measure the same three (sic) DV's:

- accuracy (whether a participant gets each item correct or incorrect)

- Reaction Time (RT). RT's will be adjusted by itemtype to only start at the start of the word that first allows participants to get the correct response. For example, if a phrase consists of Determiner Noun-with-Suffix, and the correct answer is identified by the Noun-with-Suffix, then the reaction time counter will start at 0 at the exact start of the Noun-with-Suffix. Note that this leads to the possibility of negative reaction times if a participant makes a choice before the critical part of the phrase is played auditorily.

Comprehension tests

2 Alternative Forced Choice Test (2AFC). This will test each of the untrained aliens twice, once in a way that probes grammatical number, once in a way that probes grammatical neighborhood. For both of these itemtypes, the critical word is the determiner, so no RT adjustments will be made (it will be measured from the start of the audio).

4 Alternative Forced Choice Test (4AFC). This will test each of the 30 aliens (both trained & untrained) twice, once in a way that probes for stem meaning (picture-stem assignment) and once in a way that probes for grammatical understanding. For the stem meaning trials, the critical word is the Noun-with-Suffix, so the RT will be measured from the start of that word. For the grammatical understanding trials, the critical word is the determiner, so the RT will not be adjusted (it will be measured from the start of the audio).

Error Monitoring test (EM). this test will contain grammatically correct sentences as well as sentences with three different types of errors. Aliens from the three subneighborhoods consisting of 6 aliens (small-trained, small-untrained, big-untrained) will occur in each 4 item types (gram. correct, error type 1 - determiner plurality error, error type 2 - determiner neighborhood error, error type 3 - suffix neighborhood error). In order to keep trial number as low as possible while keeping maximal power, for the big-trained subneighborhood, which consists of 12 aliens, a balanced set of 40 trials will be included in this test. This set will contain 18 error trials (6 of each type), the same number as each of the other subneighborhoods. Then, this set will contain 22 gram. correct trials (each other subneighborhood only contains 6 gram. correct trials) in order to get the total number of gram. correct trials to an acceptably high level. These 22 & 18 trials will all be balanced as well as possible between the 12 aliens, so that each alien occurs at least 3 times (some do occur 4 times) and so that errors are as well spread between these aliens as possible.

For error types 1 & 2, the determiner is the critical word, so no RT adjustments will be made for those itemtypes (it will be measured from the start of the audio). For error type 3, the Noun-with-Suffix is the critical word, so RT's for this itemtype will be measured from the start of that word. For grammatically correct items, the Noun-with-Suffix is also the critical word, since a participant has to wait to hear both words before they can know whether the full phrase is correct, so RT's for this itemtype will be measured from the start of that word.

Covariate. We will use participants' average accuracy on 4AFC noun items as a covariate in all other analyses (following Hopman & MacDonald, 2018). So e.g. if a participant gets 30/30 correct, the covariate score for this participant will be 1, if a participant gets 15/30 correct, their covariate score will be 0.5, and if a participant gets 0/30 correct, their covariate score will be 0.

Production test

For the production test, the main dependent variable will be accuracy of attempted productions (as a binary 1/0 DV), and we will also analyze the types of errors people make. Here's how the productions will be processed/scored.

Parsing. The target utterance always consists of 3 parts: determiner stem-suffix. We want to score how well a participant did on all of those 3 parts for each typed response compared to

the target response. However, in order to score the 3 parts, we first need to parse attempted responses.

Note that participants were instructed to always attempt to write any production they could come up with in the artificial language (even if it was wrong), and that if they knew a word should go there in the artificial language but couldn't think of any word, they could type the word 'something'. Thus, the word 'something' is processed slightly differently in the algorithm. We created an algorithm to code and score the attempted productions. Table F1 shows example parsings.

- if attempted production consists of one string:
    a. copy this one string into det & word2 position, then pass on to next step
- three strings:
    b. concatenate 2+3 into word2, then pass on to next step
- two strings: 1st string = det, 2nd string = noun+ suffix
    c. IF first part is the word 'something', e.g. 'somethingaaf' will be split into something (stem) + aaf (suffix)
    d. ELSE split 2nd string after first vowel cluster + consonant cluster: first part is stem, rest is suffix
- 4+ strings: throw away (score all 3 morphemes as wrong) e.g NULL NULL NULL

Table F1

Example Production Attempts and Algorithm Parsings into Determiner, Stem and Ending

production attempt	parsed determiner	parsed stem	parsed ending
lom	lom	lom	NULL
thebok	thebok	theb	ok
lom eb	lom	eb	NULL
lom thebok	lom	theb	ok
blablabla thebokoloko	blablabla	theb	okoloko
something thebok	something	theb	ok
lom somethingok	lom	something	ok
lom thebsomething	lom	theb	something
something somethingsomething	something	something	something
the aerlkinrg	the	aerlk	inrg
lom theb ok	lom	theb	ok
the pink flamingo	the	pinkfl	amingo
the crazy pink flamingo	NULL	NULL	NULL

Mapping parsed elements onto the artificial language. The three parsed elements from the attempted production (determiner, stem, ending) are compared to the target utterance.

a) Initially, we will run an algorithm using Levenshtein Distance (LD):
    - map elements user typed to closest element of correct category in language
    - or, if closest LD is equal for multiple existing elements: unmappable [written off to separate file for human inspection]
    - or, if LD to all possible morphemes of the correct category in the language is higher than a threshold: code as NULL

Threshold for an element to be marked as NULL:

- determiners: LD > 2 for LD(attempted det, existing determiner) for all four existing determiners in the language
- stems: LD > 3 for LD(attempted stem, existing stem) for all thirty existing stems in the language
- endings: LD > 2 for LD(attempted ending, existing ending) for all four existing endings in the language

b) for the unmappable elements (so LD below threshold but equal to several existing elements): These unmappable elements will be independently coded by two native English speaking Research Assistants (RAs) familiar with both the spelling and the way the artificial language sounds. The RAs will each get a simple file with a column of attempted productions that the algorithm couldn't code because one of the parsed elements had equal LD to multiple existing elements in that position. They will not have access to any other information (e.g. participant identifiers, target response, etc). For each element, they will list which existing element of the artificial language they think the participant meant. Responses from both RA's will then be compared. If both RA's list the same target element, the response will be coded as that element. If the RA's list different elements, the response will be coded as NULL.

Scoring. Once the parsed attempted productions are mapped onto either NULL or existing elements of the artificial language, they will be scored. Correct elements will be scored as 1 and incorrect (or NULL) elements will be scored as 0. The utterance as a whole will be scored as 1 if all three elements score a 1, or as 0 if one or more of the three elements was not scored as a 1.

Error Analysis. For cases where the attempted production was mappable (whether by the LD algorithm or by RA agreement) but scored as 0 (e.g. a participant wrote 'lom thebool' (an existing utterance in the artificial language) when the target utterance was 'lom thebok', we will classify which category the incorrect element belonged to and compare that to the target element. For determiners and suffixes, this means categorizing them as singular/plural and big/small neighborhood. For stems, this means categorizing them as big/small neighborhood and as trained/untrained.

Conditions

How many and which conditions will participants be assigned to?

The experiment has 2 between subjects conditions: comprehension training and production training. This determines which type of active training the participant receives. Participants will be assigned to conditions randomly by a qualtrics script, with the exception that once the total number of desired participants is reached in one condition, all participants will be assigned to the other condition so that data collection can be completed as soon as possible. While assignment is random (and thus the desired number of participants should be reached roughly equally fast in both conditions), it is possible that e.g. attrition is different in different condition, or that e.g. a technical error tied to a condition renders some data from one condition unusable.

The experiment has 2 within subjects / between items conditions:

- neighborhood; The artificial language has different determiners and suffixes for the two neighborhoods of aliens. The 'big' neighborhood consists of 18 aliens (12 trained, 6 untrained) and the small neighborhood consists of 12 aliens (6 trained, 6 untrained). Thus, in training participants see twice as many 'big' aliens, with their determiner-suffix combinations,

than 'small' aliens (while seeing each individual trained alien exactly equally often). There is a visual category distinction between the two neighborhoods, as well as the artificial language differences: 'big' neighborhood aliens have arms and legs (well, 4 limbs, they are aliens so not all have human arms/legs), giving them a somewhat humanoid and tall appearance. 'small' neighborhood aliens do not have arms/legs/limbs, giving them a more rounded appearance.

- subneighborhood (trained/untrained). 18 of the aliens appear in training (12 from the big neighborhood and 6 from the small neighborhood), these are considered 'trained' aliens. The other 12 aliens (6 from the big neighborhood and 6 from the small neighborhood) do not appear in training. After training, participants receive passive exposure to these 12 untrained aliens and their names, but no active training with their names.

Note that assignment of aliens to subneighborhoods (trained/untrained) within each neighborhood is randomized individually for each participant. So is assignment of stems to aliens within each neighborhood.

Finally, each alien appears in both singular and plural, with different suffixes and determiners within each neighborhood marking number. This is both within subjects and within items, and both training and testing are completely balanced for singular/plural.

Analyses

Comprehension tests

2AFC. Within the 2AFC test, item types (number and neighborhood) will be analyzed together but with a predictor for itemtype. Predictors: Condition*Neighborhood*Itemtype + Covariate

4AFC. Within the 4AFC test, each itemtype will be analyzed independently from the other itemtypes, as in Hopman & MacDonald 2018 (Psych Science), because each itemtype (grammar and noun trials) tests distinct knowledge. Predictors: Condition*Neighborhood*Subneighborhood + Covariate (see below for how accuracy & RT will be analyzed for all comprehension tests)

EM. The error monitoring test as a whole will be analyzed using signal detection theory, calculating a d' and bias score for each participant. These overall d' scores will be compared between the comprehension and production condition using a simple linear model. Predictors: Condition + Covariate. Within the EM test, grammatically correct and grammatically incorrect items will be independently analyzed. Predictors: Condition*Neighborhood*Subneighborhood + Covariate

Overall. Accuracy will be analyzed using generalized mixed effects regression models with the maximal random effects structure (see Barr et al. 2013 paper) that will converge. Reaction times will be analyzed using linear mixed effects regression models with the maximal random effects structure that will converge.

Production tests

Scores will be analyzed loosely following Keppenne, Hopman & Jackson (2021)'s analysis of production items. We will first analyze stem productions (on all trials) with a generalized linear mixed effects regression model. Then, for trials where participants got the stem correct, we will analyze determiner and ending accuracy (separately).

(Over)generalization errors. In addition to the analyses mentioned before, errors will be further analyzed in all tests to look at patterns of (over)generalization errors. (e.g. producing big neighborhood endings on small neighborhood stems in the production test; choosing a big neighborhood picture when a small neighborhood alien's name is played over audio in both

forced choice tests; classifying a small neighborhood stem with a big neighborhood ending in error monitoring as correct)

Outliers and Exclusions

Prescreen allows only participants who meet the following criteria to see the study in the psychology extra credit research pool:

- native language English
- not colorblind
- normal or corrected to normal vision
- normal or corrected to normal hearing

The experiment itself requires a screen size of at least 1275 x 680, we exclude participants whose screens do not meet that size (so e.g. participants can't complete the experiment on a mobile phone).

Before accessing the data, we exclude participants who:

- completed the study in the Safari browser (audio for the experiment did not work reliably in Safari)
- didn't complete any of the tests at the end of the experiment (sometimes participants would end the study early, e.g. due to personal reasons or wifi trouble)
- mention technological issues that would interfere with the content of the study in the post-experiment questionnaire (e.g. audio problems, pictures not loading, pictures and audio out of sync, etc).
- mention personal circumstances in post-experiment survey that would interfere with the content of the experiment (e.g. not being sober, having cheated on the learning experiment by taking notes, etc.)

After accessing the data, we also exclude participants who unfocused their browser during experimental content for more than 10 seconds (it's ok for participants to unfocus their browser during the built in breaks of the experiment)

For the reaction time analyses, we also exclude trials:

- that participants answered incorrectly
- with a negative adjusted RT (explained under 'analyses')
- with an RT outside of a participant's mean +- 3SD.

We do specifically analyze

- training, 2AFC & 4AFC test data for participants who finished those two tests but who ended the experiment early during the EM test.
- training, 2AFC, 4AFC & EM data, as well as all available production test items for participants who finished the comprehension tests but who either ended the experiment during the production test without finishing it, or for who certain production test items didn't load correctly (in that case, those items will be excluded, and only items where the picture did load will be analyzed).

Sample Size

We will collect participants until the end of fall semester 2021 at UW-Madison for the SONA participant pool (extra credit for intro to psychology students). This means we will be posting timeslots daily through 5 pm on December 15th, 2021. We will also be emailing participants to invite them to do the study. The participant pool has a total of ~1200 participants who meet the prerequisites this semester. Our minimum goal is to reach 100 usable participants with a complete set of data per condition.

Our goal is to run as many participants as possible before that time, with numbers as equal as possible in both conditions. Based on piloting, we are foreseeing that attrition rates might be higher in one condition (production) than the other condition (comprehension). In order to correct for that, once one condition has achieved our minimum of 100 participants, we will change the assignment to conditions. Initially that is set as 1:1, but once the goal is reached in one condition it might be set to e.g. 3:1, 2:1, etc. in order to also achieve 100 participants in the other condition while not only collection data in one condition.

We arrived at the number of a minimum of 100 participants per condition through running power simulations using pilot data. Since there are many different tests in this study (5 separately analyzed comprehension tests, for both the accuracy and RT DV's), and many different predictors (and interactions) of potential interest (7 per test including all interactions), it was not computationally feasible for us to run full power simulations (that would entail running 70 simulations, and some simulations over power curves took ~24 hours to run). The simulations we did run showed that effect sizes for different factors in different tests differed from 0.02-1.1. So, we generated a power curve for one of the most interesting factors, the three-way interaction for accuracy in the 4AFC grammar items, and it showed that 200 participants would be enough to achieve >80% power for this effect. We also generated a power curve for another very interesting factor, the three way interaction for accuracy in the grammatically incorrect EM items, and it reached >80% power around 275 participants.

Note that these power simulations took into account the higher number of comprehension than production participants that completed the pilots. Thus, the 200 participant simulation consisted of simulating 122 comprehension participants and 78 production participants, and the 275 participant simulation consisted of 161 comprehension participants and 96 production participants.

We decided that we would rather have equal numbers of participants in both conditions (or: as equal as possible), so we are aiming for a minimum of 100 participants per condition. Note that we expect power to be higher with equal numbers of participants in both conditions, so that we expect to reach 80% power for both effects of interest for which we generated power curves with those numbers.

Other

Reason for 'it's complicated' answer about has data been collected:

This study is EH's dissertation, and so there is a tight deadline for when we need to collect this data by. This study took more rounds of piloting than planned until it worked well (and until we had determined which browsers it worked well in). Once we had resolved all that, it was much later in the semester than we planned to start data collection (late October 2021; plan had been to start data collection second half of September 2021), and we knew that it would be a stretch goal to reach our desired number of 100+ participants per condition before the end of fall semester (December15th 2021). Thus, we decided to immediately start collecting data and write the preregistration while we started collecting data.

Note that while writing this preregistration, we have NOT opened participant's experimental data logs (the data we plan to analyze as registered here). We DO have access to (and have accessed this for purposes of determining how much extra credit to grant participants) the following:

- pre-experiment qualtrics survey (consent form; assesses language background, browser, screen size and randomly assigns condition; timestamp for when study is started; asks participants to conduct study alone & in quiet space)

- server timestamps for when data for each part of the experiment was sent to the server
- post-experiment qualtrics survey (timestamp for study completion; asks whether participant was alone in quiet room & did their best; asks about technological issues, e.g. audio playing ok in browser; asks if participants understood what the study was manipulating; provides debrief).

We will also use these available sources to help us determine whether a participant's data meets inclusion criteria (native English, tech issues, completed tests) and to keep track of the number of participants in each condition that meets inclusion criteria so that, if necessary, we can assign more participants to one condition once the minimum of 100 participants is reached in the other condition.

Appendix G

Full Regression Analyses

In model formulas, || indicates that all pairwise covariances between random intercept and slopes were set to 0 in order to achieve convergence and non-singularity (Barr et al., 2013). In the *p* value column of results tables, • indicates marginal significance. In all results figures, dots are model predictions, with error bars showing 95% Confidence Intervals (CIs).

Comprehension Tests

Forced Choice Tests

4AFC Stem Results. In addition to the signficant main effect for condition and the condition:familiarity interaction mentioned in the main text, I found significant main effects on accuracy for neighborhood and familiarity. Participants were significantly more accurate on stems for large neighborhood than small neighborhood aliens, and were significantly more accurate on stems for familiar than unfamiliar aliens (Table G1, Figures G1 & 7A).

Table G1

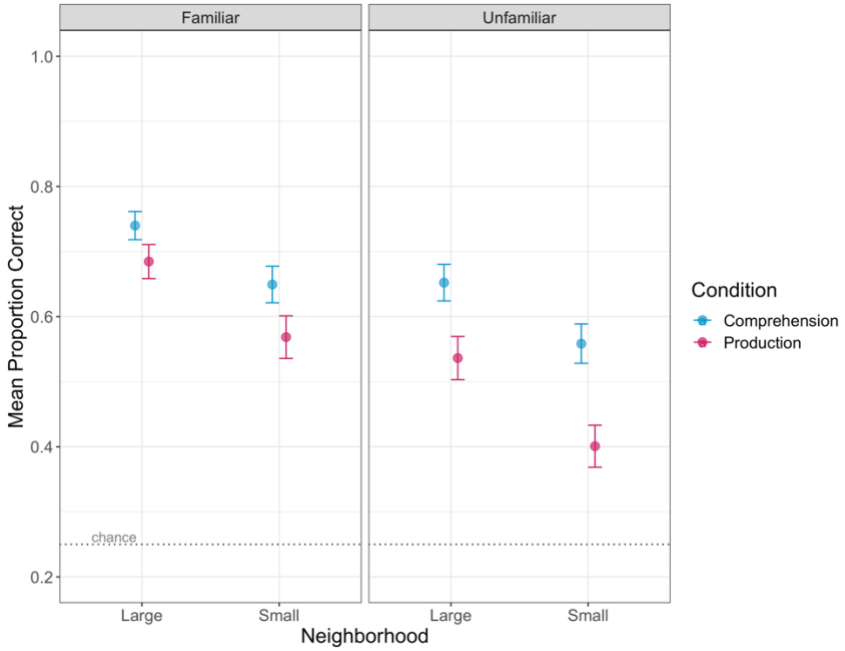
Accuracy Analysis of the 4AFC Stem Trials (depicted in Figures G1 and 7A)

Correct ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood:Familiarity||Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	0.42	0.08	5.48	< 0.001 ***
Condition	-0.43	0.15	-2.86	< 0.01 **
Neighborhood	-0.47	0.05	-8.68	< 0.001 ***
Familiarity	-0.53	0.05	-9.77	< 0.001 ***
Condition:Neighborhood	-0.11	0.11	-1.04	> 0.1
Condition:Familiarity	-0.25	0.11	-2.37	< 0.05 *
Neighborhood:Familiarity	-0.01	0.11	-0.06	> 0.1
Three-way Interaction	-0.08	0.22	-0.39	> 0.1

Figure G1

Model Predictions for the 4AFC Noun Trials Accuracy Analysis (see also Table G1)



Reaction Time analyses for these stem trials also showed significant main effects for neighborhood and familiarity: participants were not just more accurate but also faster for large than small neighborhood aliens and for familiar than unfamiliar aliens (Figure G2, Table G2).

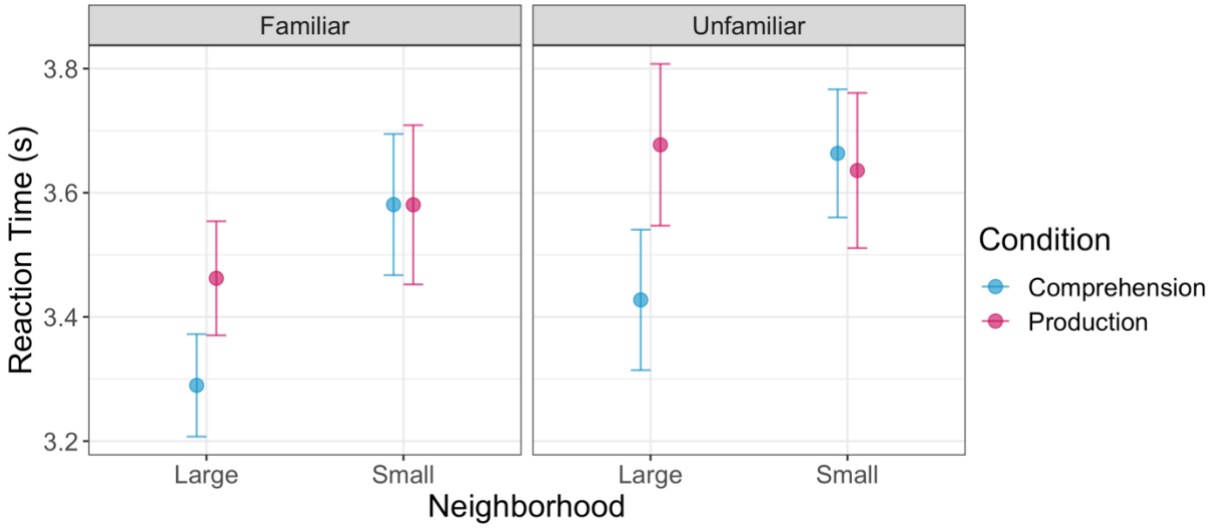
Table G2

RT Analysis of the 4AFC Noun Trials (depicted in Figure G2)

RT ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood:Familiarity|Participant) + (0+Neighborhood + Familiarity|Participant)					
	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	3.54	0.06	3357	271.2	< 0.001 ***
Condition	0.10	0.12	0.65	271.2	> 0.1
Neighborhood	0.15	0.06	6.74	267.0	< 0.001 ***
Familiarity	0.12	0.06	4.23	265.6	< 0.05 *
Condition:Neighborhood	-0.23	0.12	3.75	267.0	< 0.10 •
Condition:Familiarity	0.03	0.12	4.28	265.6	> 0.1
Neighborhood:Familiarity	-0.11	0.12	0.80	280.7	> 0.1
Three-way Interaction	-0.11	0.24	0.19	280.7	> 0.1

Figure G2

Model Predictions for the 4AFC Noun Trials RT Analysis (see also Table G2)



2AFC Results. For the Two Alternative Forced Choice trials, participants were not just more accurate (Table G3, Figure 7B), as reported in the main text, but also faster on large than small neighborhood aliens (Tables G4, Figure G3). Other than this main effect for neighborhood, no predictors reached significance for accuracy or RT.

Table G3

Accuracy Analysis of the 2AFC Task (depicted in Figure 7B)

Correct ~ LearningCondition*Neighborhood + (1+Neighborhood|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	1.35	0.08	17.50	< 0.001 ***
Condition	-0.03	0.15	-0.20	> 0.1
Neighborhood	0.18	0.08	2.36	< 0.05 *
Condition:Neighborhood	-0.03	0.13	-0.26	> 0.1

Table G4

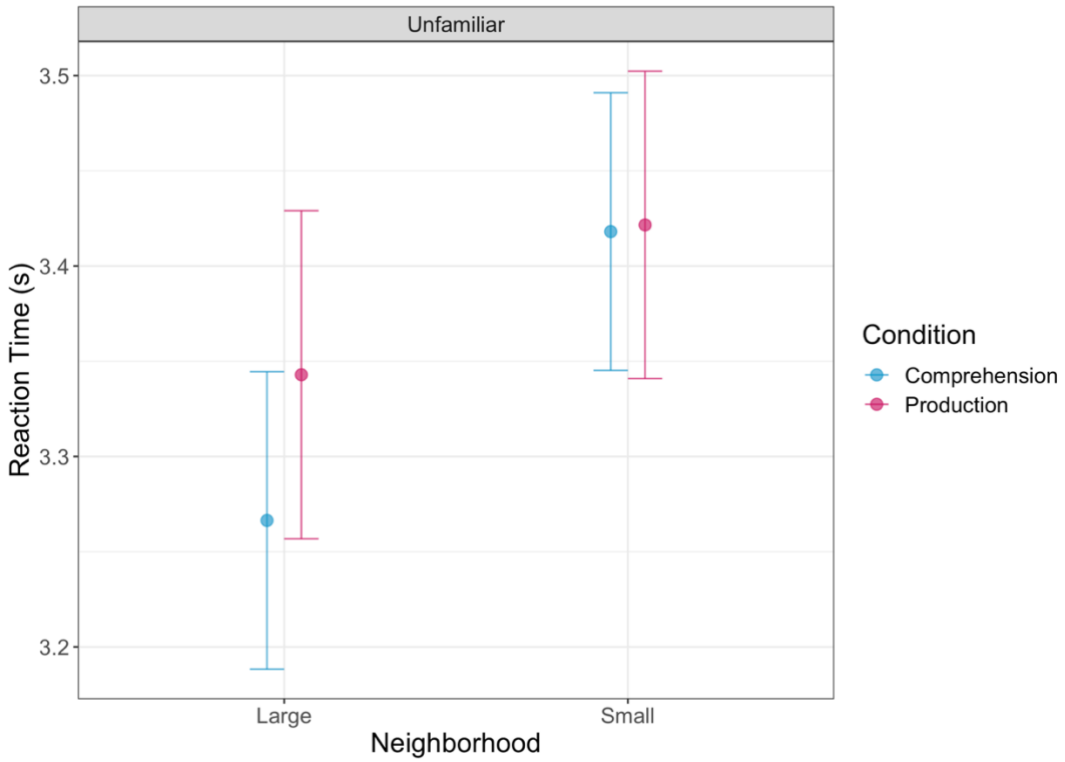
RT Analysis of the Two Alternative Forced Choice Test (depicted in Figure G3)

RT ~ LearningCondition*Neighborhood + (1+Neighborhood|Participant)

	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	3.36	0.05	4119.75	261.5	< 0.001 ***
Condition	0.04	0.11	0.15	261.5	> 0.1
Neighborhood	0.12	0.04	7.82	247.2	< 0.01 **
Condition:Neighborhood	-0.07	0.08	0.79	247.2	> 0.1

Figure G3

Model Predictions for the 2AFC Reaction Time analysis (see also Table G4)



In the methods section I indicated that I might be able to see overgeneralization errors on the neighborhood trials specifically, if participants chose large neighborhood distractors for small neighborhood targets more often than the reverse for these neighborhood trials (see Figure 5B). Thus, I conducted a separate regression analysis on the neighborhood trials to check for a significant main effect of neighborhood (with higher accuracy on large than small neighborhood aliens), which would indicate overgeneralization. However, I found no significant effects here for any predictor (Table G5).

Thus, while the 2AFC test did establish that participants in both conditions were significantly better than chance at applying the learned grammatical regularities to novel lexical items (significant intercept for accuracy), this test does not seem sensitive to condition differences. These may be ceiling results, since the 2AFC accuracy ($M = 0.75$) was higher than the mean of any other test assessing grammar comprehension in this experiment (4AFC grammar trials $M = 0.62$; EM grammatical error trials $M = 0.49$). In fact, even the numerically lowest cell mean in 2AFC (production participants on small neighborhood aliens: $M = 0.74$) was higher than the highest cell means in both the 4AFC grammar trials (comprehension participants on the large neighborhood familiar aliens: $M = 0.68$) and the EM tests (production participants on suffix neighborhood errors for large neighborhood familiar aliens: $M = 0.68$).

When I initially designed this experiment, it did not include the 4 blocks of passive trials to introduce the unfamiliar (then called test-only) aliens. While the two alternative forced choice results in that version were more interesting and showed potential condition differences, pilot testing showed that error monitoring and even 4AFC grammar tests had at or below chance performance. Production test performance as well as post-experiment debriefs showed that it was virtually impossible for participants to complete the production trial. Pilot participants expressed frustration about having to produce items they had not learned, and indicated that this frustration affected even performance on items they had learned (familiar aliens). Including the 4 blocks of passive trials showed an immediate improvement in piloting results for the 4AFC, EM and PT. I did not realize at the time that this would also push 2AFC performance to ceiling.

Table G5

Accuracy Analysis of the 2AFC Neighborhood Items (Second Round)

Correct ~ LearningCondition*Neighborhood + (1|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	1.41	0.09	16.49	< 0.001 ***
Condition	-0.08	0.17	-0.49	> 0.1
Neighborhood	0.12	0.09	1.36	> 0.1
Condition:Neighborhood	-0.14	0.18	-0.77	> 0.1

4AFC Grammar results. In addition to the effects mentioned in the main text, I found a significant main effect of Familiarity on accuracy. Participants were more accurate on familiar than unfamiliar aliens (Table G6, Figure G4). This effect was in the expected direction, it makes sense that participants are more accurate on trained, familiar aliens than untrained, unfamiliar aliens.

Table G6

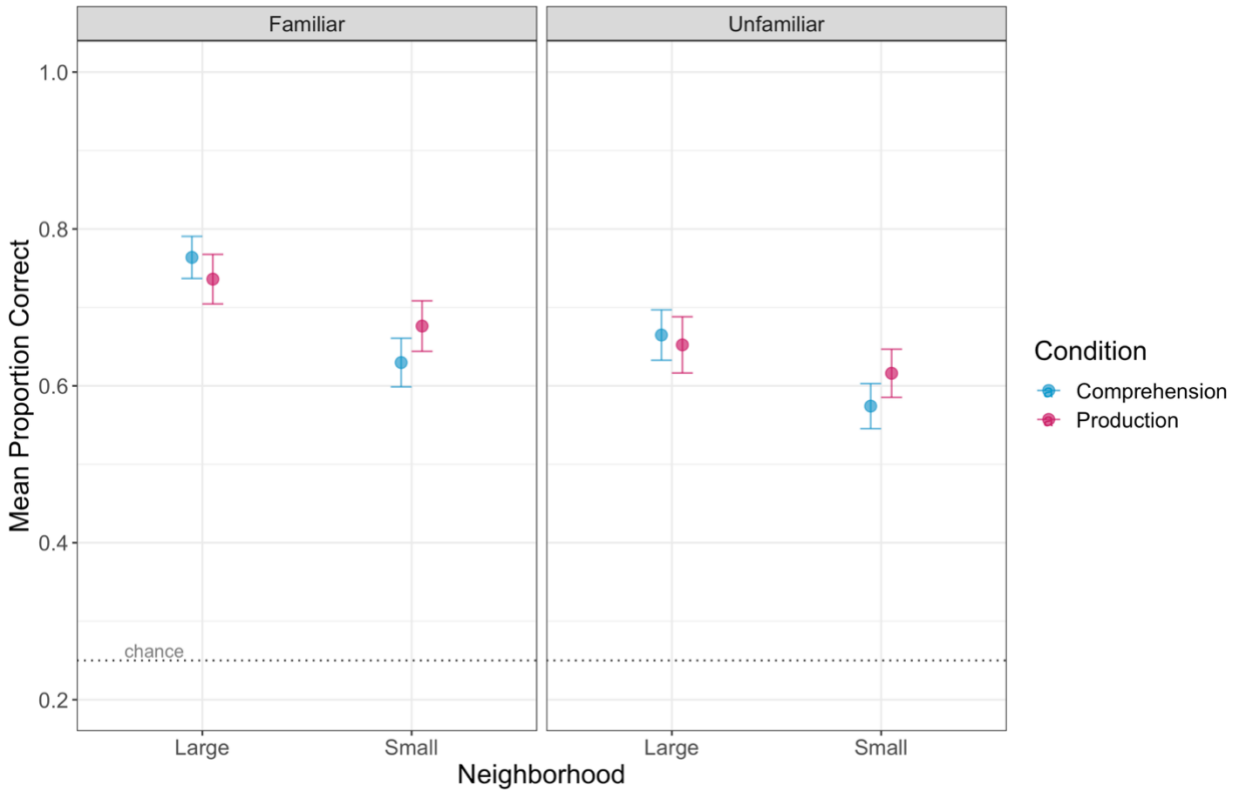
Accuracy Analysis of the 4AFC Grammar Trials (depicted in Figures 7B and G4)

Correct ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood + Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	0.69	0.09	8.00	< 0.001 ***
Condition	0.04	0.17	0.26	> 0.1
Neighborhood	-0.37	0.07	-5.31	< 0.001 ***
Familiarity	-0.35	0.06	-5.59	< 0.001 ***
Condition:Neighborhood	0.29	0.13	2.22	< 0.05 *
Condition:Familiarity	0.03	0.12	0.26	> 0.1
Neighborhood:Familiarity	0.20	0.11	1.77	< 0.1 •
Three-way Interaction	-0.12	0.22	-0.56	> 0.1

Figure G4

Model Predictions for the 4AFC Grammar Trials Accuracy Analysis (see also Table G6)



In addition to the main effect for neighborhood on accuracy mentioned in the main text, participants are also significantly faster on trials with large than small neighborhood aliens (Table G7, Figure G5). This effect was in the expected direction, it makes sense that participants are both faster and more accurate on the more common large than the small neighborhood aliens. Participants are thus overall better at grammar understanding on large than small neighborhood aliens.

For both accuracy and RT, there was also a marginal neighborhood by familiarity interaction. In both cases, the neighborhood effects were smaller for for unfamiliar than familiar aliens. I did not predict or expect this, but it seems reasonable that neighborhood differences in grammatical understanding come out more clearly for trained than untrained items.

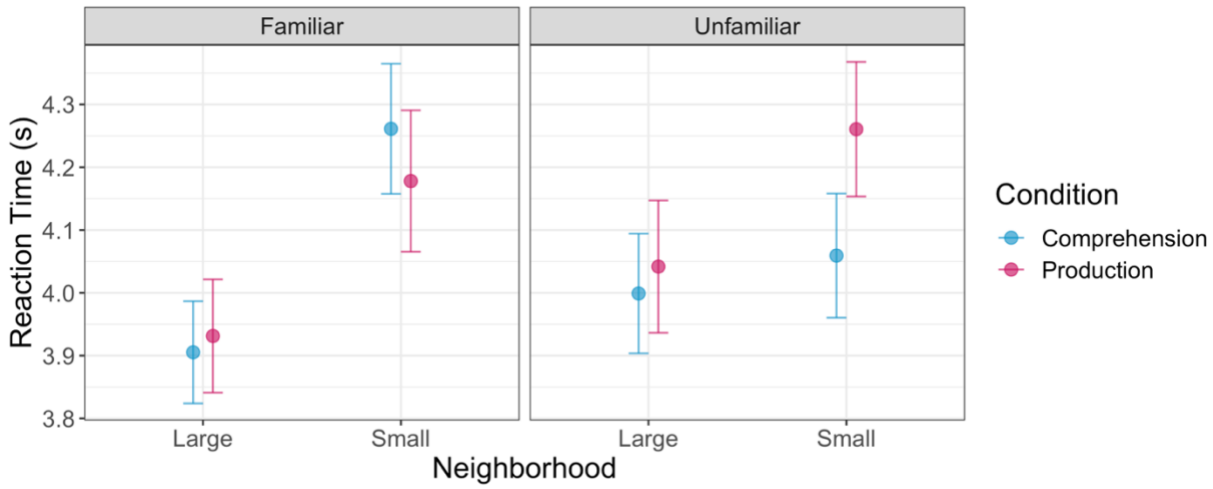
Table G7

RT Analysis of the 4AFC Grammar Trials (depicted in Figure G5)

RT ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood:Familiarity|Participant) + (0+Neighborhood + Familiarity|Participant)					
	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	4.08	0.06	5235	261.8	< 0.001 ***
Condition	0.05	0.11	0.17	261.8	> 0.1
Neighborhood	0.22	0.05	17.14	241.5	< 0.001 ***
Familiarity	0.02	0.05	0.18	252.1	> 0.1
Condition:Neighborhood	0.03	0.11	0.05	241.5	> 0.1
Condition:Familiarity	0.15	0.10	2.35	252.1	> 0.1
Neighborhood:Familiarity	-0.16	0.09	3.07	236.1	< 0.1 •
Three-way Interaction	0.27	0.18	2.09	236.1	> 0.1

Figure G5

Model Predictions for the 4AFC Grammar Trials RT Analysis (see also Table G7)



In the methods section I indicated that I might also be able to see overgeneralization errors by looking at which foils participants chose in these trials when they made errors. Of course, as explained in the methods section, in these 4AFC trials choosing the correct neighborhood and choosing the correct alien are equivalent. In order to analyze this, I coded whether participants chose the correct neighborhood (so the correct alien in either plurality) or not (the distractor alien in either plurality). Analyzing this, I only found an effect of familiarity on likelihood of choosing the correct neighborhood. Participants were better at choosing the correct neighborhood for familiar than unfamiliar aliens (Table G8, Figure G6). This effect is expected: it stands to reason that it was easier for participants to choose the correct neighborhood (or alien) for familiar than unfamiliar aliens. However, I did not find more evidence of overgeneralization errors here, since there were no effects of neighborhood.

Table G8

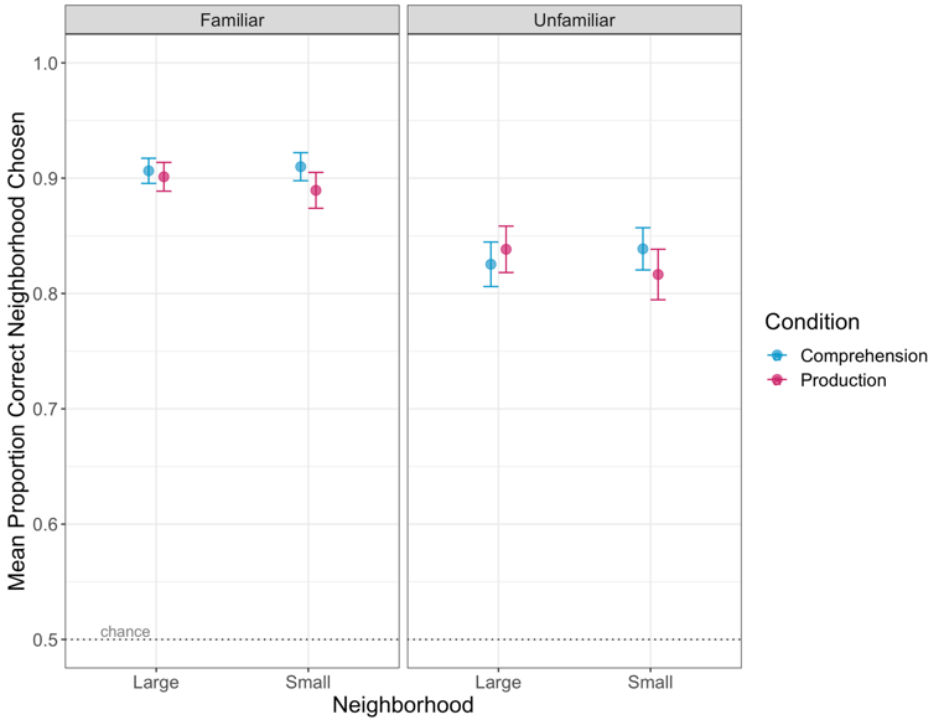
Correct Neighborhood Chosen Analysis results for 4AFC Grammar Trials (depicted in Figure G6)

Correct_Neighborhood ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood + Familiarity||Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	1.90	0.08	23.25	< 0.001 ***
Condition	-0.09	0.16	-0.56	> 0.1
Neighborhood	-0.04	0.07	-0.53	> 0.1
Familiarity	-0.64	0.08	-8.04	< 0.001 ***
Condition:Neighborhood	-0.21	0.13	-1.57	> 0.1
Condition:Familiarity	0.11	0.16	0.72	> 0.1
Neighborhood:Familiarity	0.01	0.13	0.10	> 0.1
Three-way Interaction	-0.08	0.26	-0.31	> 0.1

Figure G6

Model Predictions for 4AFC Grammar Trials Correct Neighborhood Chosen Analysis (Table G8)



Error Monitoring Test

Main analyses. As preregistered, in addition to the overall signal detection theory d' analysis reported in the main text, I analyzed each error type separately.

Wrong Plurality Determiner Errors (Type 1). In addition to the signficant main effect for condition on accuracy mentioned in the main text, participants were significantly more accurate and faster at catching these errors on large than small neighborhood aliens (Tables G9-10, Figures G6-7). These effects were in the expected direction: catching errors should be easier for the large than the small neighborhood.

Participants were also significantly more accurate and marginally faster at catching these wrong plurality determiner errors for unfamiliar than familiar aliens. This is unexpected: I had predicted that any type of grammatical error would have been easier to catch on familiar than unfamiliar aliens.

Table G9

Accuracy Analysis of EM Wrong Plurality Determiner Errors (depicted in Figure G6)

Correct ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood + Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	0.09	0.10	0.91	> 0.1
Condition	0.54	0.20	2.66	< 0.01 **
Neighborhood	-0.51	0.10	-5.05	< 0.001 ***
Familiarity	0.32	0.07	4.72	< 0.001 ***
Condition:Neighborhood	-0.23	0.20	-1.16	> 0.1
Condition:Familiarity	-0.03	0.13	-0.26	> 0.1
Neighborhood:Familiarity	0.16	0.13	1.22	> 0.1
Three-way Interaction	0.20	0.25	0.78	> 0.1

Figure G6

Model Predictions for the EM Wrong Plurality Determiner Errors Accuracy Analysis (Table G9)

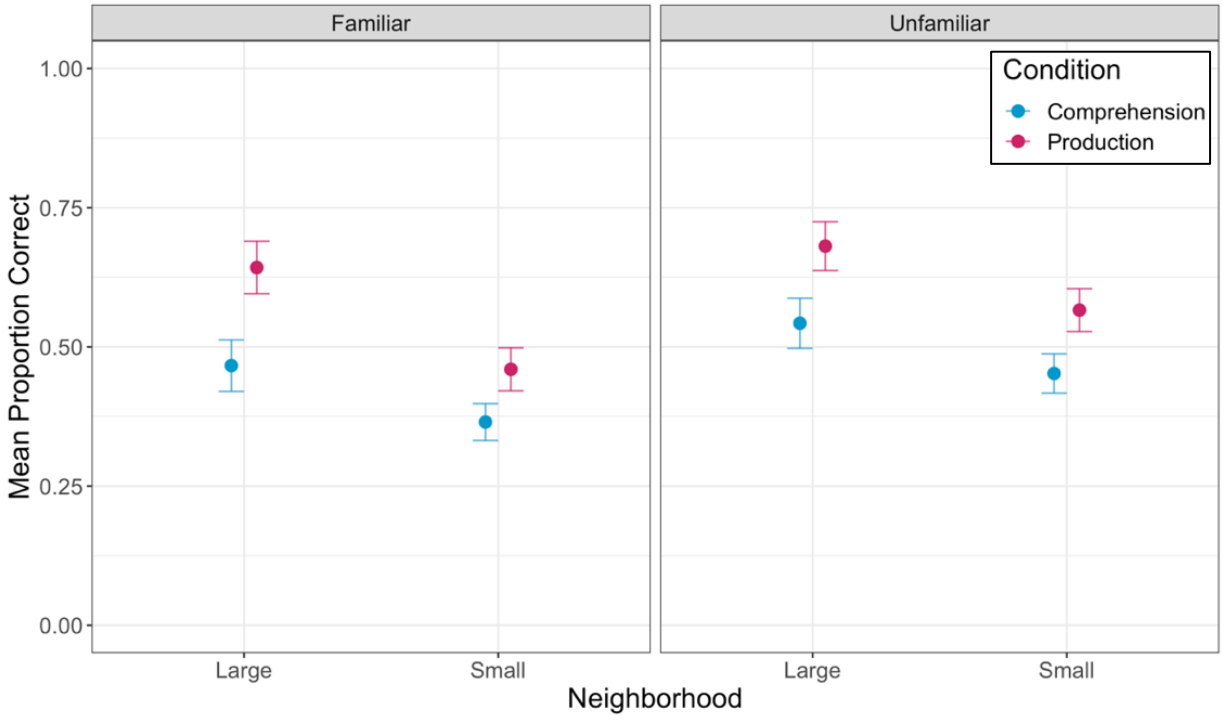


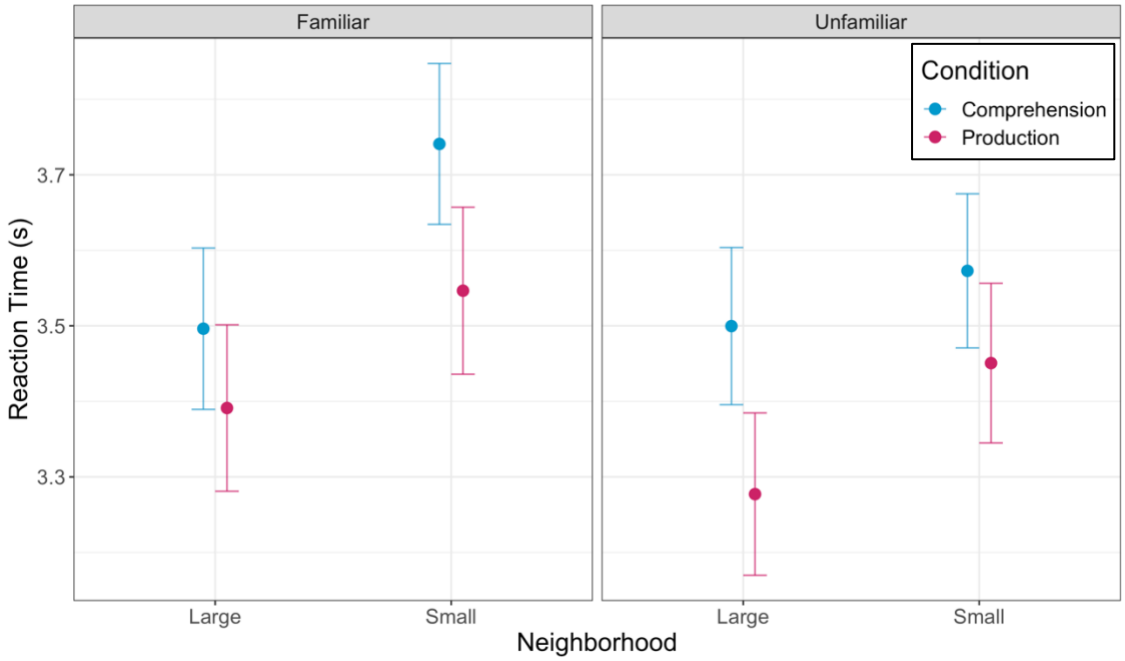
Table G10

RT Analysis of the EM Wrong Plurality Determiner Errors (depicted in Figure G7)

RT ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood|Participant)					
	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	3.50	0.06	3458	224.8	< 0.001 ***
Condition	-0.16	0.12	1.83	224.8	> 0.1
Neighborhood	0.16	0.06	7.32	184.4	< 0.01 **
Familiarity	-0.09	0.05	3.28	2439.1	< 0.1 •
Condition:Neighborhood	0.01	0.12	0.00	184.4	> 0.1
Condition:Familiarity	-0.02	0.10	0.05	2439.1	> 0.1
Neighborhood:Familiarity	-0.08	0.10	0.55	2439.1	> 0.1
Three-way Interaction	0.19	0.21	0.85	2439.1	> 0.1

Figure G7

Model Predictions for the EM Wrong Plurality Determiner Errors RT Analysis (see also Table G10)



Wrong Neighborhood Determiner Errors (Type 2). Production participants were not just significantly more accurate but also marginally faster at detecting these errors than comprehension participants (Table G12, Figure G9). This corroborates that production participants were all around better at catching determiner neighborhood errors than comprehension participants.

Participants were also significantly more accurate at catching these errors for small than large neighborhood aliens (Table G11, Figure G8). This is unexpected: I had predicted any type of grammatical error to be easier to catch on large than small neighborhood aliens.

Furthermore, participants were also significantly faster and significantly more accurate at catching these wrong neighborhood determiner errors for unfamiliar than familiar aliens. This is unexpected: I had predicted that any type of grammatical error would have been easier to catch on familiar than unfamiliar aliens.

Table G11

Accuracy Analysis of EM Wrong Neighborhood Determiner Errors (depicted in Figure G8)

Correct ~ LearningCondition*Neighborhood*Familiarity + (1+Familiarity|Participant) + (1+Neighborhood + Neighborhood:Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.22	0.10	-2.39	< 0.05 *
Condition	0.80	0.19	4.22	< 0.001 ***
Neighborhood	0.34	0.09	3.90	< 0.001 ***
Familiarity	0.20	0.07	2.86	< 0.01 **
Condition:Neighborhood	-0.14	0.18	-0.77	> 0.1
Condition:Familiarity	0.06	0.14	0.42	> 0.1
Neighborhood:Familiarity	0.05	0.13	0.37	> 0.1
Three-way Interaction	0.31	0.26	1.18	> 0.1

Figure G8

Model Predictions for the EM Wrong Neighborhood Determiner Accuracy Analysis (Table G11)

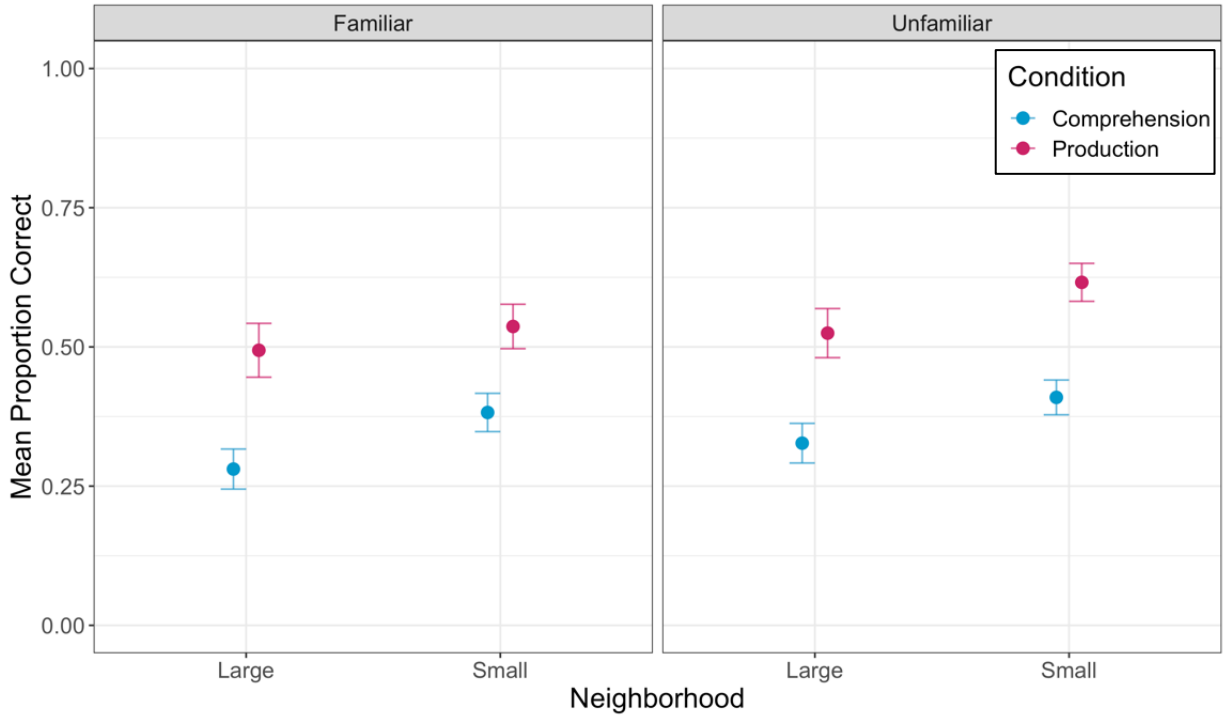


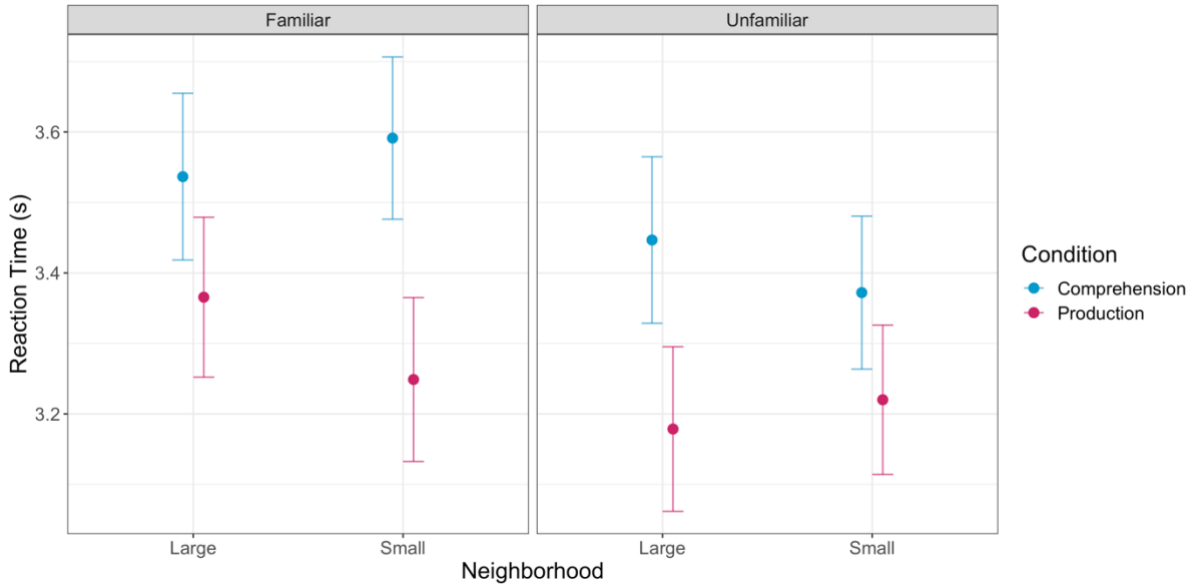
Table G12

RT Analysis of EM Wrong Neighborhood Determiner Errors (depicted in Figure G9)

RT ~ LearningCondition*Neighborhood*Familiarity + (0+Familiarity|Participant) + (1+Neighborhood:Familiarity|Participant)					
	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	3.37	0.06	3170	219.3	< 0.001 ***
Condition	-0.23	0.12	3.80	219.3	< 0.1 •
Neighborhood	-0.02	0.06	0.16	2204.8	> 0.1
Familiarity	-0.13	0.06	4.52	170.9	< 0.05 *
Condition:Neighborhood	-0.03	0.12	0.05	2204.8	> 0.1
Condition:Familiarity	0.05	0.12	0.14	170.9	> 0.1
Neighborhood:Familiarity	0.02	0.13	0.01	183.6	> 0.1
Three-way Interaction	0.29	0.27	1.13	183.6	> 0.1

Figure G9

Model Predictions for the EM Wrong Neighborhood Determiner Errors RT Analysis (Table G12)



Wrong Neighborhood Suffix Errors (Type 3). In addition to the signficant main effect for Condition on accuracy mentioned in the main text, participants were significantly more accurate at catching these errors on large than small neighborhood aliens (Tables G13, Figure G10). This effect was in the expected direction: catching errors should be easier for the large than the small neighborhood.

Participants were also more accurate at catching these errors for familiar than unfamiliar aliens. This effect was in the expected direction: catching errors should be easier for the familiar than for unfamiliar aliens.

Furthermore, there was a significant condition by familiarity interaction for these wrong neighborhood suffix errors. The condition difference was larger for familiar than unfamiliar aliens. I did not predict or expect this, but it seems reasonable that condition differences in grammatical understanding come out more clearly for trained than untrained items.

Table G13

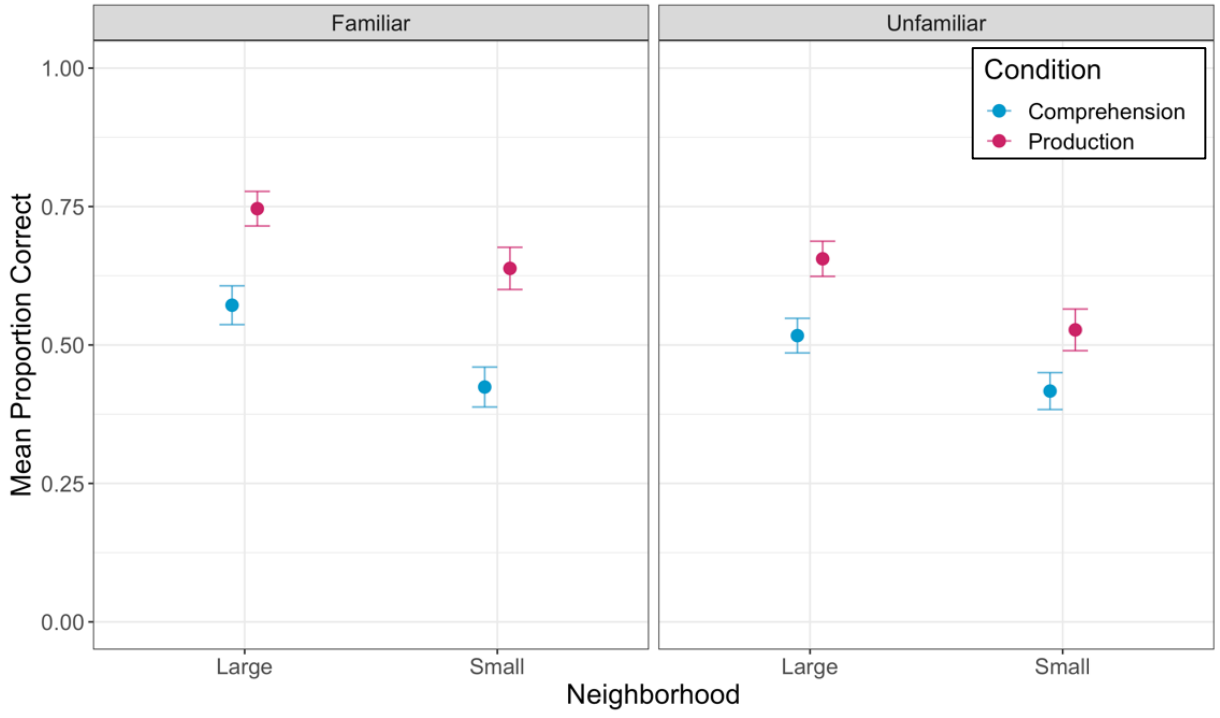
Accuracy Analysis of EM Wrong Neighborhood Suffix Errors (depicted in Figure G10)

Correct ~ LearningCondition*Neighborhood*Familiarity + (1+Familiarity|Participant) + (1+Neighborhood + Neighborhood:Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	0.26	0.08	3.15	< 0.01 **
Condition	0.67	0.17	4.00	< 0.001 ***
Neighborhood	-0.51	0.08	-6.39	< 0.001 ***
Familiarity	-0.29	0.07	-4.07	< 0.001 ***
Condition:Neighborhood	-0.03	0.16	-0.14	> 0.1
Condition:Familiarity	-0.32	0.14	-2.30	< 0.05 *
Neighborhood:Familiarity	0.08	0.13	0.66	> 0.1
Three-way Interaction	-0.22	0.25	-0.86	> 0.1

Figure G10

Model Predictions for the EM Wrong Neighborhood Suffix Errors Accuracy Analysis (Table G13)



There were no significant or marginal effects in the RT analysis for these wrong neighborhood suffix errors (Table G14, Figure G11).

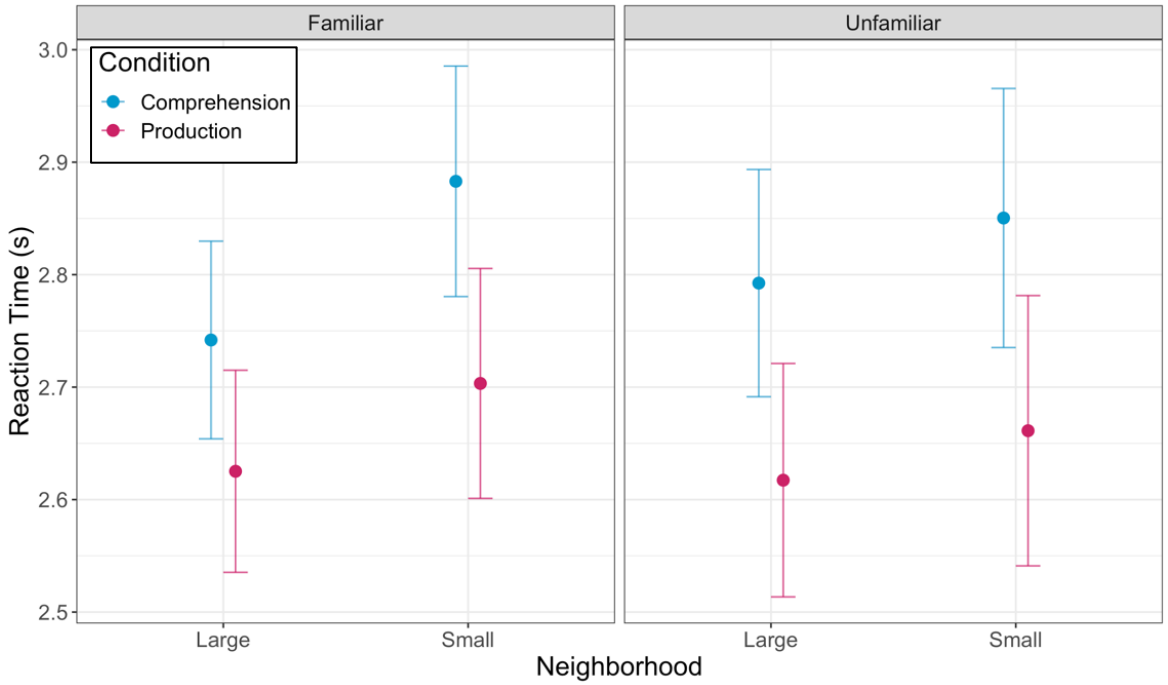
Table G14

RT Analysis of EM Wrong Neighborhood Suffix Errors (depicted in Figure G11)

RT ~ LearningCondition*Neighborhood*Familiarity + (1+Neighborhood + Familiarity|Participant)					
	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	2.73	0.06	2389	224.3	< 0.001 ***
Condition	-0.17	0.11	2.19	224.3	> 0.1
Neighborhood	0.08	0.06	2.13	192.2	> 0.1
Familiarity	-0.01	0.06	0.02	203.9	> 0.1
Condition:Neighborhood	-0.04	0.11	0.12	192.2	> 0.1
Condition:Familiarity	-0.03	0.12	0.08	203.9	> 0.1
Neighborhood:Familiarity	-0.06	0.10	0.33	2648.8	> 0.1
Three-way Interaction	0.05	0.20	0.06	2648.8	> 0.1

Figure G11

Model Predictions for the EM Wrong Neighborhood Suffix RT Analysis (see also Table G14)



Post-hoc exploratory analyses. There were three unexpected main effects: (1) for wrong neighborhood determiner errors (type 2), participants were more accurate on small than large neighborhood aliens and (2-3) for both types of wrong determiner errors (types 1 & 2, wrong plurality determiner and wrong neighborhood determiner), participants were more accurate on unfamiliar than familiar aliens. In both cases, the error type(s) of interest were compared with wrong neighborhood suffix trials (type 3). For (1) this comparison to only error type 3 made sense as a comparison because both error types 2 and 3 are neighborhood errors, with the only difference the morpheme that contained the error. For (2-3), both error types 1 and 2 were determiner errors, and thus they were compared to error type 3, with a different error location, namely the suffix. Since these were post-hoc exploratory follow-up analyses, only these specific effects of interest were interpreted, although full glmer's are reported in Tables A.20-21.

In order to analyze (1), the novel predictor error location was created: error type 2, wrong neighborhood determiner errors, were coded as -0.5, and error type 3, wrong neighborhood suffix errors, as 0.5. Data for those two error types were then analyzed together (Table G15).

In addition to the stem neighborhood by error location interaction of interest that is mentioned in the main text, there is also an error location by familiarity interaction, caused by the unexpected finding that participants are more accurate at catching wrong neighborhood determiner errors for unfamiliar than familiar aliens (3). Since, as mentioned at the start of this section a similar main effect of familiarity in favor of unfamiliar aliens was also present for the wrong plurality determiner errors (2), a separate follow-up analysis with all three error types together was conducted.

Table G15

Neighborhood Error Serial Order Analysis (depicted in Figure 9)

Correct ~ LearningCondition*Neighborhood*Familiarity*Location + (1+Familiarity|Participant) + (1+Neighborhood:Location + Neighborhood:Familiarity + Familiarity*Location |Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	0.05	0.08	0.54	> 0.1
Condition	0.76	0.17	4.51	< 0.001 ***
Neighborhood	-0.08	0.05	-1.71	< 0.1 •
Familiarity	-0.04	0.05	-0.73	> 0.1
Error Location	0.50	0.08	6.34	< 0.001 ***
Condition:Neighborhood	-0.06	0.09	-0.66	> 0.1
Condition:Familiarity	-0.10	0.11	-0.94	> 0.1
Neighborhood:Familiarity	0.08	0.10	0.82	> 0.1
Condition:Location	-0.14	0.16	-0.86	> 0.1
Neighborhood:Location	-0.84	0.14	-5.86	< 0.001 ***
Familiarity:Location	-0.47	0.10	-4.83	< 0.001 ***
Cond.:Neigh.:Fam.	0.08	0.19	0.40	> 0.1
Cond.:Neigh.:Loc.	0.09	0.29	0.31	> 0.1
Cond.:Fam.:Loc.	-0.38	0.19	-1.96	< 0.1 •
Neigh.:Fam.:Loc.	0.03	0.18	0.16	> 0.1
4-way interaction	-0.53	0.36	-1.50	> 0.1

Note. Cond.: Condition, Neigh.: Neighborhood, Fam.:Familiarity, Location & Loc.:Error Location

In order to analyze a factor with three levels (the three error types), two orthogonal contrasts were created. The first contrast was coded to capture error location (wrong plurality determiner: 0.33, wrong neighborhood determiner: 0.33, wrong neighborhood suffix: -0.67). The second contrast, error type, captured whether there was a difference between the two different types of determiner errors (wrong plurality determiner: -0.5, wrong neighborhood determiner: 0.5, wrong neighborhood suffix: 0). In a contrast analysis, a pattern is considered confirmed if the contrast of interest (in this case, the error location by familiarity interaction) is significant, and the remaining orthogonal contrast is not. To control for type I error rate in this exploratory, multiple contrast analysis, p-values for the contrast of interest were Scheffé-corrected (error location (c1) / error Type (c2) by familiarity interactions). There was indeed a significant error location by familiarity interaction, and no error type by familiarity interaction (Table G16, Figure 10). Thus, the contrast analysis confirmed the pattern I had spotted.

Table G16

Error Location by Familiarity Contrast Analysis (depicted in Figure 10)

Correct ~ LearningCondition*Familiarity*(Location+Type) + (1+Familiarity + Location + Type|Participant)					
	Coeff.	Std. Error	*z* value	*p* value	*p* Scheffé
Intercept	0.07	0.08	0.93	> 0.1	-
Condition	0.67	0.15	4.42	< 0.001 ***	-
Familiarity	0.08	0.04	1.70	< 0.1 •	-
ErrorLocation (contrast 1)	-0.35	0.07	-5.36	< 0.001 ***	-
ErrorType (contrast 2)	-0.25	0.11	-2.34	< 0.05	-
Condition:Familiarity	-0.05	0.09	-0.52	> 0.1	-
Condition:Location (c1)	-0.02	0.13	-0.16	> 0.1	-
Condition:Type (c2)	0.26	0.22	1.23	> 0.1	-
Familiarity:Location (c1)	0.49	0.08	6.48	< 0.001 ***	< 0.001 ***
Familiarity:Type (c2)	-0.13	0.09	-1.44	> 0.1	> 0.1
Cond.:Fam.:Location (c1)	0.32	0.15	2.11	< 0.001 ***	-
Cond.:Fam.:Type (c2)	0.16	0.18	0.88	> 0.1	-

Note. Coeff.: Coefficient; Std. Error: Standard Error; Cond.: Condition, Fam.:Familiarity, Location & Loc.:Error Location (contrast 1), Type: Error Type (contrast 2). Neighborhood was left out of this analysis to ensure convergence, because it was not a relevant predictor in this analysis.

Production Test

Overall Accuracy

Overall accuracy was analyzed separately for each morpheme (Tables G17-19). Besides the effects reported in the main text, there was a significant main effect of neighborhood for all three morphemes, so that participants more accurately produced determiners, stems and suffixes for large than the small neighborhood aliens. This effect was in the expected direction: it makes sense that it was easier to correctly produce morphemes describing large neighborhood than small neighborhood aliens.

There was also a significant main effect of familiarity for all three morphemes, so that participants more accurately produced determiners, stems and suffixes for familiar than for unfamiliar aliens. This effect was also in the expected direction: it makes sense that it was easier to correctly produce morphemes describing familiar than unfamiliar aliens.

Finally, for suffixes only, there was a significant neighborhood by familiarity interaction, so that the main effect for neighborhood was smaller for unfamiliar than familiar aliens. I did not predict or expect this, but it seems reasonable that neighborhood differences in producing suffixes come out more clearly for trained than untrained items. Note that this effect is parallel to neighborhood by familiarity interactions for the 4AFC grammar trials.

Table G17

Accuracy Analysis of PT Determiner Productions (depicted in Figure 11A)

Correct ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood*Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.35	0.13	-2.76	< 0.01 **
Condition	1.33	0.25	5.25	< 0.001 ***
Neighborhood	-0.76	0.13	-5.94	< 0.001 ***
Familiarity	-0.36	0.05	-6.60	< 0.001 ***
Condition:Neighborhood	0.37	0.26	1.42	> 0.1
Condition:Familiarity	-0.12	0.11	-1.06	> 0.1
Neighborhood:Familiarity	0.03	0.10	0.26	> 0.1
Three-way Interaction	0.20	0.21	0.95	> 0.1

Table G18

Accuracy Analysis of PT Stem Productions (depicted in Figure 11B)

Correct ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood*Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-1.52	0.10	-14.67	< 0.001 ***
Condition	0.05	0.21	0.24	> 0.1
Neighborhood	-0.50	0.08	-6.58	< 0.001 ***
Familiarity	-1.23	0.10	-11.96	< 0.001 ***
Condition:Neighborhood	-0.27	0.13	-2.03	< 0.05 *
Condition:Familiarity	-0.69	0.19	-3.59	< 0.001 ***
Neighborhood:Familiarity	0.01	0.15	0.06	> 0.1
Three-way Interaction	-0.51	0.26	-1.92	< 0.1 •

Table G19

Accuracy Analysis of PT Suffix Productions (depicted in Figure 11C)

Correct ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood + Neighborhood:Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.62	0.10	-6.21	< 0.001 ***
Condition	0.53	0.20	2.65	< 0.01 **
Neighborhood	-1.13	0.09	-13.20	< 0.001 ***
Familiarity	-0.66	0.05	-14.00	< 0.001 ***
Condition:Neighborhood	0.27	0.17	1.56	> 0.1
Condition:Familiarity	-0.12	0.09	-1.25	> 0.1
Neighborhood:Familiarity	-0.20	0.10	-2.11	< 0.05
Three-way Interaction	-0.03	0.19	-0.18	> 0.1

Overgeneralization

Model for both neighborhoods. For all three morphemes, there was a main effect of neighborhood, with the proportion of neighborhood errors always larger for small than for large neighborhood aliens (Tables G20-22; Figure G12). These neighborhood errors on small aliens are of particular interest, because they constitute overgeneralizations. Thus, as expected, the production test elicited many overgeneralization errors.

There was also a main effect of familiarity for all three morphemes, with the proportion of neighborhood errors always larger for familiar than for unfamiliar aliens. This was as expected: participants made more overgeneralization errors for unfamiliar than for familiar aliens.

Then, there was a significant main effect of condition for determiners and suffixes (both grammatical morphemes), but not for stems. Comprehension participants made significantly more neighborhood errors than production participants on determiners and suffixes. Thus, comprehension participants made more overgeneralization errors on grammatical morphemes than production participants.

For determiners, there was a significant condition by familiarity interaction. The condition difference was larger for familiar than unfamiliar aliens.

Table G20

Analysis of Proportion Neighborhood Errors for Identifiable Determiner Productions (Figure G12A)

Neighborhood Error ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood*Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-1.46	0.11	-13.87	< 0.001 ***
Condition	-0.78	0.21	-3.74	< 0.001 ***
Neighborhood	1.07	0.21	5.22	< 0.001 ***
Familiarity	0.60	0.08	7.35	< 0.001 ***
Condition:Neighborhood	-0.47	0.40	-1.16	> 0.1
Condition:Familiarity	0.37	0.15	2.57	< 0.05 *
Neighborhood:Familiarity	-0.21	0.16	-1.34	> 0.1
Three-way Interaction	-0.02	0.29	-0.08	> 0.1

Table G21

Analysis of Proportion Neighborhood Errors for Identifiable Stem Productions (Figure G12B)

Neighborhood Error ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood*Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-1.26	0.07	-18.93	< 0.001 ***
Condition	-0.06	0.13	-0.42	> 0.1
Neighborhood	0.81	0.11	7.74	< 0.001 ***
Familiarity	0.55	0.09	6.51	< 0.001 ***
Condition:Neighborhood	-0.13	0.20	-0.66	> 0.1
Condition:Familiarity	0.22	0.16	1.39	> 0.1
Neighborhood:Familiarity	-0.09	0.16	-0.55	> 0.1
Three-way Interaction	0.35	0.31	1.15	> 0.1

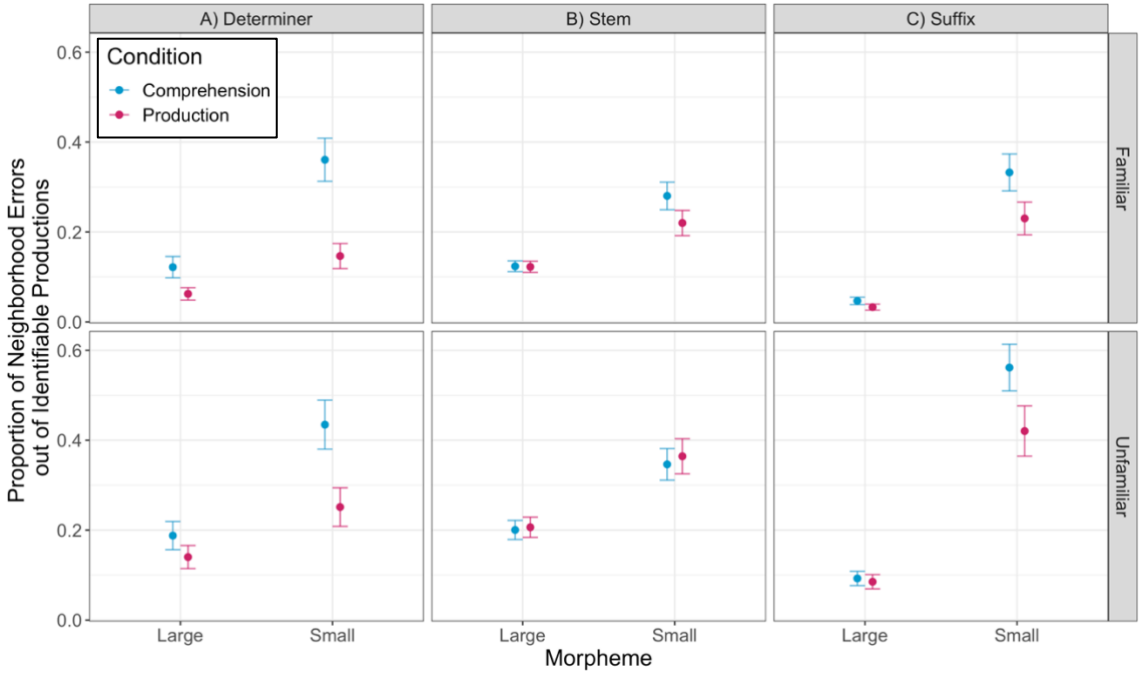
Table G22

Analysis of Proportion Neighborhood Errors for Identifiable Suffix Productions (Figure G12C)

Neighborhood Error ~ Condition*Neighborhood*Familiarity + (1+ Neighborhood*Familiarity|Participant)				
	Coefficient	Standard Error	*z* value	*p* value
Intercept	-1.63	0.09	-17.46	< 0.001 ***
Condition	-0.39	0.18	-2.18	< 0.05 *
Neighborhood	2.27	0.18	12.42	< 0.001 ***
Familiarity	0.89	0.10	8.62	< 0.001 ***
Condition:Neighborhood	-0.31	0.34	-0.90	> 0.1
Condition:Familiarity	0.11	0.17	0.66	> 0.1
Neighborhood:Familiarity	0.05	0.21	0.21	> 0.1
Three-way Interaction	-0.33	0.34	-0.99	> 0.1

Figure G12

Overgeneralizations in the Production Test..



Simple Effects for Small Neighborhood. As explained in the main text, neighborhood errors were of particular interest when they occured for the small neighborhood, since those neighborhood errors constituted overgeneralizations. Thus, the same model presented in Tables A.26-28 was rerun centered on the small neighborhood (small_neighborhood: large = -1, small = 0). Mathematically, this means that for the other predictors, instead of providing the overall condition (or familiarity, etc.) effect, the predictor now provided the estimate for condition (or familiarity, etc.) specifically as it held for the small neighborhood aliens. Thus, an effect of condition (or familiarity, etc.) in this analysis held specificially for the overgeneralization errors of interest (Tables A.29-31, Figure 11). Note that results for the four predictors that include small_neighborhood (small_neighborhood, condition:small_neighborhood, xmall_neighborhood:familiarity and the three-way interaction) were identical to those printed in Tables G20-G22 and were thus left out of Tables G23-G25.

Table G23

Analysis of Proportion Neighborhood Errors for Identifiable Determiner Productions Centered on Small Neighborhood (Figure 12A)

Neighborhood Error ~ Condition*SmallNeighborhood*Familiarity + (1+ SmallNeighborhood*Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.92	0.15	-6.27	< 0.001 ***
Condition	-1.01	0.29	-3.45	< 0.001 ***
Familiarity	0.49	0.11	4.61	< 0.001 ***
Condition:Familiarity	0.36	0.20	1.79	< 0.1 •

Table G24

Analysis of Proportion Neighborhood Errors for Identifiable Stem Productions Centered on Small Neighborhood (Figure 12B)

Neighborhood Error ~ Condition*SmallNeighborhood*Familiarity + (1+ SmallNeighborhood*Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.85	0.10	-8.87	< 0.001 ***
Condition	-0.12	0.19	-0.64	> 0.1
Familiarity	0.51	0.12	4.14	< 0.001 ***
Condition:Familiarity	0.40	0.24	1.67	< 0.1 •

Table G25

Analysis of Proportion Neighborhood Errors for Identifiable Suffix Productions Centered on Small Neighborhood (Figure 12C)

Neighborhood Error ~ Condition*SmallNeighborhood*Familiarity + (1+ SmallNeighborhood*Familiarity|Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	-0.50	0.14	-3.65	< 0.001 ***
Condition	-0.54	0.27	-2.00	< 0.05 *
Familiarity	0.92	0.12	7.76	< 0.001 ***
Condition:Familiarity	-0.06	0.23	-0.25	> 0.1

Appendix H

Dankwoord

To my first ever research mentor, Dr. Caitlin Fausey. I am lucky that you took a chance on a random student emailing you about re-using some stimuli. I would never have even thought to apply for PhD programs in the USA (especially not in the Midwest) if it hadn't been for the honors project I got to do at IU-Bloomington as a master's student. Being mentored by you showed me what I wanted (and still want) to be like as a mentor myself. Hearing your voice to ask me the first zoom question during my defense was a beautiful full circle moment for me.

Tiffany and the Seppa family, thank you for helping me get settled in Madison that very first year. (Grampa) Dan, your driving lessons will stick with me for a lifetime.

Teresa, you were my first ever mentee. I remember going to the PIF conference in Leuven together, and it's still surreal to me that you lived in the Netherlands for a year as a master's student. Levi, I doubt that my job interview at Duolingo would've gone as well as it did without your help practicing live-coding. Miriam, sometimes I worried we are too alike and would drive each other crazy with our perfectionism and work ethic, but we made it to graduation! Jur, jij bent mijn allernieuwste en allerlaatste mentee, and I absolutely love having gotten this opportunity to work together on research. To all the mentees in between, thank you for teaching me all you did. Mentoring y'all has been a privilege, and the letters of recommendation that Teresa, Jamie, Charles, Emily and Robin wrote meant more than the mentoring award itself did.

Shout-out to Polly, Taryn & Jennifer, for helping me stay sane during graduate school.

To the department staff, thanks for taking care of everything so that we can focus on research. John, thank you for helping me coordinate such a smooth hybrid format dissertation defense. It meant the world to me to get to have my friends & family from the Netherlands, as well as colleagues spread over the US, attend my dissertation and even ask questions live. Kevin, I'm sure you're happy to see me go ;) of course I'm joking, but wow did we have many administrative hiccups with my appointments, visa and benefits over the years. I always felt a lot more secure knowing that you had my back!

Ken, Kristin, Pippa & (little) Olive; Mike, Olive & Lance; Kathy, Denny & Mickey; Kathleen, Melissa & Charlie; Emily & Ryder; Ben, Murphy & Sasha; Emily, Turk & Jager; Margaret, Izzie & family; Scott, Hippo, Blossom & Lily; Susan & Mindo; thank you all for tolerating the buoyant fluffy little ball of anxiety that was an adolescent Vinya. The 7 am weekday mornings by the woods were always a bright way to start my day, and particularly helped me cope with isolation during the start of the pandemic.

Snowflower Sangha, thank you for welcoming me and providing me with a spiritual home. Amanda, Amy, Beng, Celeste, Curt, Elizabeth, Gloria & Walt, Jon, Karuna & Micha, Leah & Zach & Linnea, Lester, Lisa, Mary & David, Nan & Finn, Nancy & Joe, Rosebud, Sarah, Sherrie & Doug, Steven, Susan & Jim, Tod, Tom, Tony, and everyone else I've interacted with. Wishing you all mettā, karuṇā, muditā, upekkhā in abundance.

Mark, Martin, Matt CB, Arella, Desia, Steve & Cassie, thanks for putting up with me as your chattiest office-mate and teaching me all you did in the process, whether it was about complicated analyses or about a random English word I couldn't come up with while writing. Y'all encouraged me to have confidence in my growing abilities as a researcher, and to keep being my quirky, exuberant self in the process. Alyssa, Anna, Chris, Emily, Gaylen, Jeff, Kim, Kushin, Lilia, Matt B, Melissa, Michelle, Misty, Mitch, Odile, Phil, Pierce, Ron, Sarah, Sasha, Sean, Viri, my time in the department was brighter because of you all.

To all of my friends & family who watched my dissertation defense online, thank you for joining! It made me so happy to see all of your names and faces right in front of me as I was presenting my thesis.

Janneke, Suus, Michiel & Astrid, I know I tend to pop in and out, maar als we elkaar spreken voelt het als vanouds, and I miss living closer to you all and seeing more of each other.

Joe, I'm grateful for your co-mentorship during my middle years in grad school. I've really enjoyed working with you, your ability to think outside the box stands out in particular, as well as your ability to make people working in seemingly disparate fields of research into a cohesive lab! Our bilingualism and creativity paper is the most efficient division of labor for getting a project done and a paper out that I've ever gotten to be part of. I'm grateful to you and Maryellen both for letting me flexibly move in and out of Joe's lab as made sense for me, I really appreciate that freedom & flexibility to get to do what's felt right at different times for me and my career.

Steve, having you officiate Jay and my marriage ceremony in our backyard in Madison at the start of the pandemic is one of my most special memories as a grad student. I can't wait for your wedding with Courtney, and hope we'll keep visiting each other as our careers progress.

Carolina, te extrañaré en Pittsburgh. ¡Y gracias por todo que has hecho por nosotros!

Maryellen, I remember how star-struck I was when we met at the Valkhof museum in Nijmegen at the small comprehension=production workshop and I got to have dinner with you. Such a contrast with a recent phonecall where you couldn't even understand me because I was panicking about my dissertation defense potentially needing to be rescheduled. Thank you for supporting me through the good times and the bad. I really appreciate your thoughtful guidance, and your ability to keep me focused without curbing my enthusiasm. I've worked with plenty of intelligent people, but your brilliance is unmatched.

Maaike[2] & Tim. Waar te beginnen? Chocofonduen in de tentamenweek en samen Love Story meeblèren, Marieweekenden, studiereizen, de Efteling. Vriendjes zien komen, gaan, en blijven. Maaike G, ik weet nog dat je me meenam naar dat scoutingfeestje, en die bbq waar ik de enige vegetariër was. Ik heb nog steeds dat jurkje ('de sok') in de kast hangen dat wij ooit samen in Arnhem kochten. Maaike Z, 'ik ben niet zo goed in talen leren' en 'Elise 'bitch' Hopman'; glæder mig til at besøge dig og Henrik i Danmark ;). Way to go getting the post-doc grant & taking your research career to the next level. Tim, schat, dawno, dawno temu była dziewczyna, która chciała się nauczyć języka polskiego, en dat nu voor Duolingo gaat werken. Even after all of these years we continue to walk in step, niet alleen onze zelfde propedeuse & bachelor tegelijk af, maar zelfs onze verschillende masters & PhD's tegelijk afgerond. Ik weet nog dat we in Portland proostten op het feit dat mijn eerste artikel voor review was geaccepteerd, en van de zomer in Pittsburgh kunnen we mooi vieren dat we beiden onze dissertation afgerond hebben.

Kat, Chels & David, PhDivas, cheerleaders and my closest friends here in Madison. David & Kat, we all lost our dads during grad school, and now we're all graduating during the same semester. Chels & David, it means so much to me that I got to show you Nijmegen. Kat & Chels, we started together and became friends quickly in our first year. David, your quiet genius and grace are qualities I can only ever strive for. Kat, I've always looked up to you as a researcher and a person, and will probably keep doing so forever. Chels, from nerding out about personal finance to teaching me how to put on make-up for a job interview, there's nothing too small or too large for us to share. All of you were there in our backyard when we got married at the start of the pandemic, and the kindle you got me that summer was the most meaningful present I've ever received. I hope these friendships are for life.

Kris, Alma & Kaly, I always feel at home with you all. Kris, I've known you for more than half of my life; wie had kunnen bedenken dat een vriendschap die begon met een paar geleende boeken would still be going strong by now? And what a lovely surprise to have you ask the final question at my dissertation defense! Alma, mi hermana mexicana, espero que algún día vivamos cerca la una de la otra. Kaly, ik word er altijd blij van om jou weer te zien!

Jill, Jeff, Jordan, Jarad & Lowen, being far away from my own family in the Netherlands I feel especially lucky to have loving family on this side of the ocean. Jill, your support and wisdom have helped me make this country home. Lowen, I can't express how special it's been for me to witness your first six months of life, and how excited I am to see you grow up.

Mamma, Meike, Jannes & Noura, het betekent zò veel voor me dat jullie in levende lijve bij mijn verdediging en uitreiking konden zijn (en Mike via Zoom!), en dat we daarna met zijn allen naar Zion National Park zijn geweest, de eerste gezinsvakantie in heel veel jaar. Jannes, ik kan niet wachten om ook bij jouw verdediging te zijn over een paar jaar. Meike, ik hoop dat we na Yosemite en nu Zion nog veel national parks samen gaan verkennen. Mam, we missen pappa allemaal nog steeds, en dat maakt me extra dankbaar voor jouw goede gezondheid, en dat jij mijn trouwen, mijn promotie en hopelijk nog heel veel andere mijlpalen kan meemaken.

Jay, Plato & Sophia love you (at least) as much as they do me, which is kind of unfair since they've known me twice as long, but I guess that's cats for you. Adopting Vinya was probably the scariest thing we ever did together, and her boundless, unconditional love touches me (and you, and everyone else who she can reach with her enthusiastic face-licks) daily. When I moved here from the Netherlands to get my PhD at UW-Madison I hadn't really planned to permanently immigrate to the USA. I know you likewise hadn't planned to leave your downtown high rise apartment for a house with a yard on the west side, or to leave Madison to move to Pittsburgh. Let's keep changing each other's plans for a long time to come.