For learning morphological rules, production tasks are no better than comprehension

Elizabeth Pankratz*, Aislinn Keogh*, Simon Kirby, Jennifer Culbertson
Centre for Language Evolution, University of Edinburgh
**Joint first author*

e.c.pankratz@sms.ed.ac.uk
aislinn.keogh@ed.ac.uk

Abstract

This paper weaves together two strands of previous research: one which identifies that adults struggle to learn morphological rules, and another which indicates that language learning can be facilitated by language production. Here, we ask: can a production task help adults learn morphological rules? In two artificial language learning experiments, we taught participants a language that indicated thematic role with both a fixed word order (a word-level rule, which should be easier for adults to learn) and case marking (a morphological rule, which should be harder for adults to learn). We manipulated whether participants practised this artificial language using a comprehension task or a production task, and then asked whether participants who did the production task were more likely to learn the case marking rule. We also assessed how aware participants were of the morphological pattern that results from the case marking, even if they did not associate certain markers with certain thematic roles *per se*. Experiment 1 tested L1 English participants, and Experiment 2 tested L1 German participants: populations that differ in their prior experience of case. In both experiments, we found that participants across the board failed to learn the case marking rule, even though the majority did detect the morphological pattern that was the consequence of case marking. We conclude that the production task we used in this study did not suffice to help adults learn a less accessible morphological rule.

Keywords: artificial language learning, case marking, word order, language production, segmentation

1 Introduction

Learners of a new language will often discover that no matter how many target-language books, films, or podcasts they absorb, their language skills do not truly blossom until they have practised producing the language themselves.

Language production benefits both infant learners of their first language as well as adult learners of further languages. In L1 acquisition, children who use their target language more frequently show stronger expressive abilities in that language throughout development, independent of their level of comprehension (Bohman et al., 2010; Donnelly and Kidd, 2021; Ribot et al., 2018). And in L2 acquisition, production tasks have been shown to improve how L1 Mandarin users learn English relative clauses (Izumi,

2002) and how people with diverse L1s learn German grammatical gender (Keppenne et al., 2021); the way production tasks benefit L2 acquisition has been influentially referred to as the Output Hypothesis (Swain, 2005). Artificial language learning studies also illustrate that production practice helps adults both to learn rules (Hopman and MacDonald, 2018) and to generalise them (Hopman, 2022).

A separate strand of research has shown that adult learners don't acquire all kinds of rules equally well. Particularly troublesome are morphological rules; adult learners' difficulty with both nominal and verbal inflectional morphology has been well documented (see, e.g., Bentz and Winter, 2013; Holmes and Dejean De La Bâtie, 1999; Parodi et al., 2004; Rogers, 1987; Sagarra and Ellis, 2013). Case marking poses a particular challenge: L2 learners of German and Turkish struggle to learn the case-marking morphology, even if their L1 also has case (Jordens et al., 1989; Papadopoulou et al., 2011). In contrast, rules that apply to larger chunks, such as words and phrases, seem more accessible. For example, when learning noun classification systems, adults tend to rely more on class membership cues that do not require them to segment below word level (i.e., determiners) compared to sub-word cues that do require segmentation (i.e., suffixes; Keogh and Lupyan, in press). And typological evidence also suggests that adults may prefer word-level rules: languages with more adult L2 learners tend to be morphologically simpler (Bentz and Winter, 2013; Lupyan and Dale, 2010).

In this study, we ask whether a production task can cause adult learners to acquire a hard-to-learn morphological rule that requires words to be segmented, moving beyond an easier word-level rule that requires no segmentation. This question builds on intriguing results from Hopman and MacDonald (2018). In their artificial language learning experiment, participants who did a production task seem to have acquired morphological rules better than word-level ones. Specifically, those participants appear to be more sensitive to errors in suffixing than errors in word order (see their Figure 5, p. 968).

However, this boost to morphological rule learning is just a descriptive result that the original paper does not explore further. Additionally, this finding might come not from the production task *per se*, but rather from properties of the artificial language that Hopman and MacDonald used. The language was very complex in that every sentence contained multiple modifiers and adverbial phrases, so the word order rules might have been hard to identify. On the other hand, several words in every sentence contained identical suffixes, making the morphological pattern highly salient. Here, we aim to follow up on Hopman and MacDonald's result using an artificial language designed to test whether production tasks help adults learn morphological rules over word-level rules.

Why would we expect language production to help adults learn morphological rules at all? One explanation for why production strengthens language learning is what Swain (2005) describes as its "noticing role". The idea is that when people produce a language, they process it more deeply and are thus more likely to notice linguistic patterns and induce possible generalisations (see also Izumi, 2002). As long as the task is more active than a recognition-based comprehension task, we would expect the noticing role of production to take effect; based on literature on the effects of different kinds of tests, any kind of test beyond passive recognition should improve learning (see, e.g., Kang et al., 2007; McDaniel et al., 2007; McDermott et al., 2014). We therefore hypothesised that a production task could draw people's attention to morphological patterns that they may otherwise have missed.

As a testing ground for this hypothesis, we used the well-studied trade-off between case marking, an example of a morphological rule, and fixed word order, an example of a word-level rule (Bentz and Winter, 2013; Fedzechkina et al., 2011; Levshina, 2020; Lupyan and Dale, 2010). The rest of this paper discusses two preregistered experiments (`https://osf.io/qbjda/?view_only=cdd6600d8e9c45e5bc153058ea97df29`) that test this hypothesis on two populations that differ in their prior experience with case-marking systems: Experiment 1 tests L1 English participants, and Experiment 2 tests L1 German participants.

To foreshadow our results: overall, participants learned the fixed word order rule but failed to acquire the case marking rule, although the majority did notice the recurring syllable pattern that was the consequence of case marking. Even participants already familiar with the concept of case (the German L1 participants in Experiment 2) showed the same clear preference to treat words as the smallest unit in the language and not to segment below this level. With respect to our main hypothesis, we found that taking part in a production task does not make participants more likely to learn the case marking rule. This result suggests that, although production tasks may generally facilitate learning, they don't necessarily help adult learners to discover morphological rules.

2 Experiment 1

Participants were trained on a series of sentences that each described a transitive event between two human characters. These sentences were designed to be compatible with both word-level and morphological strategies for marking thematic role. Specifically, each sentence had the same fixed word order (SOV), a consistent word-level cue, *and* each noun bore a suffix corresponding to its grammatical role (nominative for the agent role and accusative for the patient role), a consistent morphological cue.

For example, participants might see an image of a fairy pushing a doctor and learn the corresponding sentence *fuvu zijo gix*. Then they might see a cowboy kicking a pirate and learn the sentence *lovu wujo kuv*. In both sentences, the word order is SOV, and in both sentences, the agent is marked with -*vu* and the patient with -*jo*. Thus participants could analyse the language in two different ways: like (1), in which nouns remain unsegmented, or like (2), in which nouns are segmented into stem and case marker.

(1)	a.	fuvu	zijo	gix	
		fairy	doctor	push	
	b.	lovu	wujo	kuv	
		cowboy	pirate	kick	
(2)	a.	fu-vu	zi-jo	gix	
		fairy-NOM	doctor-ACC	push	
	b.	lo-vu	wu-jo	kuv	
		cowboy-NOM	pirate-ACC	kick	

Note that the recurring syllables at the end of each noun could also be analysed in terms of their linear order: participants could arrive at an analysis like "the first noun always ends in *vu*, and the second noun always ends in *jo*". This is not a case marking analysis *per se*, since it's not based on thematic roles. But it is still of interest to us, because we're concerned with how well participants can identify patterns below word level. For this reason, in what follows, we refer to the two possible analyses not as "fixed word order" and "case marking", but rather as "unsegmented" and "segmented", respectively.

A crucial aspect of the training phase's design is that participants received no direct evidence that nouns have morphological structure, because none of the characters appeared as both agent and patient. Thus, the language's grammar is ambiguous. To illustrate concretely: a participant would only ever see the fairy character as an agent, only ever labelled as *fuvu*. They receive no information about what form this word would take if the fairy were a patient. The word might become *fujo*, following the segmented analysis, or remain *fuvu*, following the unsegmented analysis. Thus it was possible for participants to successfully learn the training data without segmenting the words.

After training, we split participants into two groups to introduce the manipulation by task. Half of the participants practised the sentences they had learned using a more active production task, while the other half practised using a more passive comprehension task. The production task involved constructing sentences by clicking on the component syllables in the correct order, while the comprehension task simply involved choosing the correct image from an array of two.

Next, in the testing phase, we showed participants the same scenes they saw in training, but with the characters' roles reversed. For example, where in training they saw a fairy pushing a doctor, now they saw the doctor pushing the fairy. We then asked them to judge two different sentences that might describe this scene. The first kind of sentence was formed using the unsegmented analysis: the full words for the agent and patient were rearranged. The second kind of sentence was formed using the segmented analysis: only the stems were rearranged, and the case markers stayed in place.

If participants learned the nouns as unsegmented, holistic chunks, they should accept the first kind of sentence. If they segmented the nouns into stem and suffix, they should accept the second kind of sentence. Given previous findings that adults struggle to learn case morphology, we expected our participants to prefer sentences formed using the unsegmented analysis. However, crucially, here we test whether this preference is affected by the type of practice task they did: comprehension or production.

Finally, participants completed a one-shot cloze task with a novel character, i.e., a character held out from the set encountered in training. The goal here was to assess whether participants were aware of the language's morphological patterns (i.e., that the first noun always ends in a particular syllable, and that the second noun always ends in another), whether or not they actually analysed these syllables as case markers.

2.1 Materials

The artificial language contained transitive sentences made up of three words: one for the agent, one for the patient, and one for the action, in that order (i.e., SOV). All verbs were monosyllabic CVC nonsense words, and all nouns were disyllabic CVCV nonsense words. Verbs were randomly selected from a set of 28: *gax*, *gix*, *gox*, *hix*, *jeg*, *jix*, *juf*, *juz*, *kex*, *kez*, *kuv*, *kux*, *nuz*, *puv*, *pux*, *vaf*, *vof*, *wez*, *wox*, *zax*, *zok*, *zox*, *zud*, *zuf*, *zug*, *zup*, *zuv*, and *zux*. Nouns were randomly assembled from nine possible stem syllables (*bu*, *fu*, *gu*, *ki*, *lo*, *ru*, *wu*, *ze*, and *zi*) and two suffix syllables (*vu* and *jo*) such that all agent nouns took one suffix and all patient nouns took the other.

Each sentence accompanied an image, a line drawing of two human characters interacting. A few examples are shown in Figure 1. The nine possible characters were: a chef, a cowboy, a doctor, a fairy, a footballer, a nun, a pirate, a princess, and a wizard. Each scene showed the agent character engaging in a reversible transitive action toward the patient character. The nine possible actions were: admiring, greeting, kicking, kissing, patting, poking, pushing, seeing, and yelling. Each image had two mirrored versions: one with the agent on the left, and one with the agent on the right.

To keep the artificial lexicon easily learnable, we randomly selected only six characters and two actions for each participant. The characters and actions were randomly associated with nonsense noun stems and verbs from the sets listed above. Then, each character was mapped to the thematic role they would appear in during the training phase. The mapping between characters and roles was random, with one constraint: we disallowed any permutations in which all agents were female and all patients were male (or vice versa), to forestall analyses of the suffixes as gender markers. All in all, participants saw 18 unique scenes during training: 3 agents × 3 patients × 2 verbs.

2.2 Procedure

We wrote the experiment in JavaScript using the jsPsych library (de Leeuw et al., 2023). It contains four phases, detailed below and illustrated in Figure 1.

2.2.1 Training

In each training trial, participants saw an image alone for 1000 ms. Then the corresponding sentence in the artificial language appeared below it. 2500 ms later, a 'next' button appeared below the sentence. Clicking on it advanced participants to the next trial.

The whole training phase consisted of three blocks of 18 trials each, one trial per scene. Participants could optionally take a short break between each block.

2.2.2 Practice

After training, participants were divided into two groups: one group completed a production practice task (the PRODUCTION condition), and the other completed a comprehension practice task (the COMPREHENSION condition). Both practice tasks involved familiar scenes and sentences that participants had encountered during training.

Participants in the PRODUCTION condition saw a familiar scene and had to build the correct sentence for this scene out of its component syllables. Below the image were five gaps, and below the gaps was one button per syllable in the sentence, shown in a random order. Although this task is less active than, say, speaking the artificial language sentence aloud, it still involves reproducing the linguistic signal that participants received. This reproduction places additional demands on participants that the comprehension task, as a simple recognition task, does not (more detail on the comprehension task below).

Clicking one of the buttons added that syllable into the leftmost gap in the sentence, so the sentence was filled in left to right as each syllable was clicked. An 'undo' button

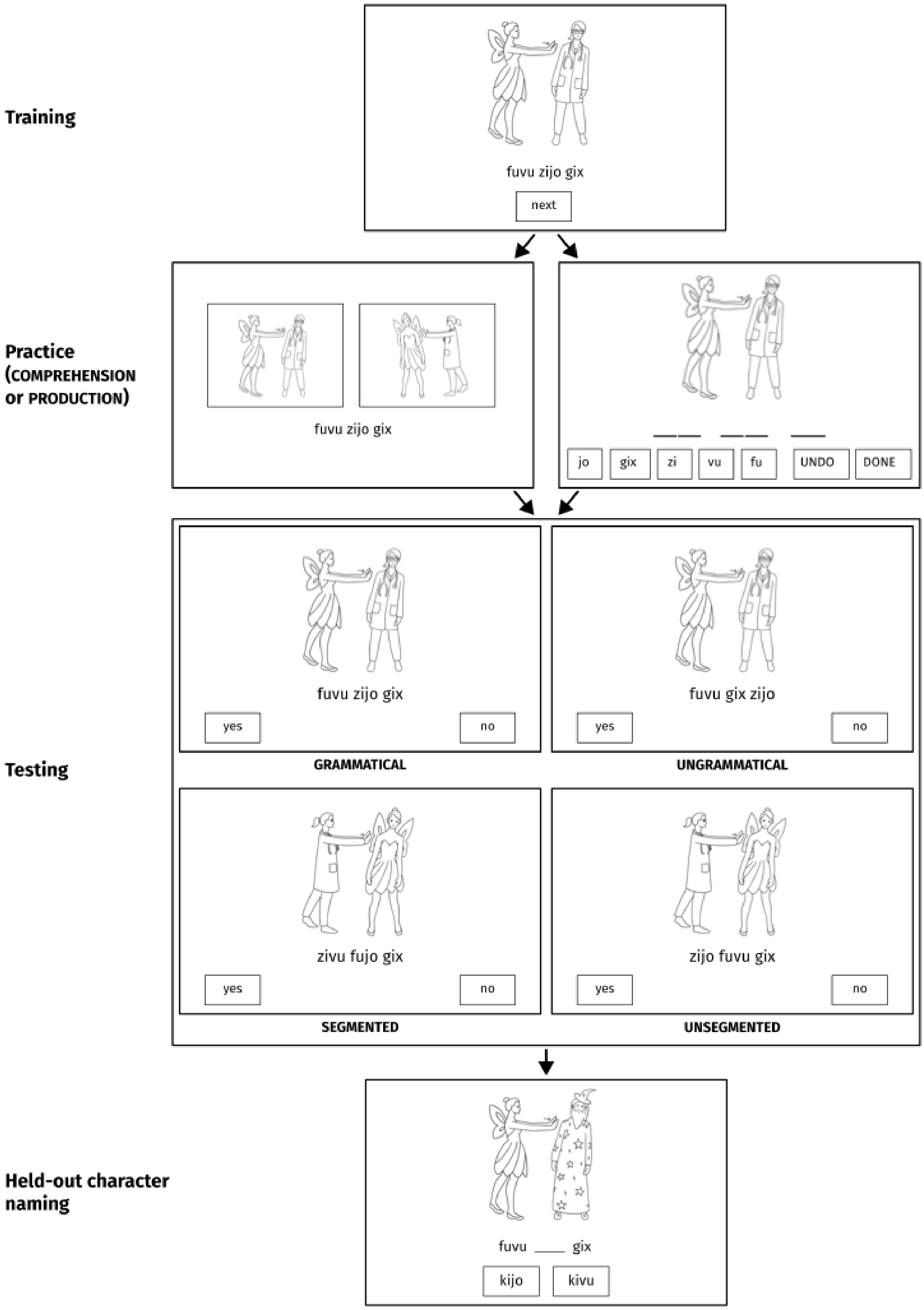

Figure 1: A schematic overview of the four phases of the experiment. All participants do the same training, then do either the comprehension or the production practice task. Then all participants complete the same testing and character naming phases. In other words, the two conditions differ *only* in the practice task.

emptied the most recently filled gap. Participants could submit their sentence with the 'done' button as long as the sentence included every syllable once.

After submitting the sentence, participants received feedback on their response and were shown the correct sentence. The feedback stayed on-screen until participants clicked 'next' to continue. Each participant did 18 production trials, one per familiar scene, shown in a random order.

Participants in the COMPREHENSION condition were shown a familiar sentence and had to select the corresponding scene from an array of two. The target scene was a familiar one encountered during training; the foil image contained the same characters but with the thematic roles reversed (that is, if the target showed the fairy pushing the doctor, then the foil would show the doctor pushing the fairy). The order of target and foil was randomised on each trial. The agent appeared on the left in one image and on the right in the other, so that the characters themselves remained in the same position in each image.

So that we do not confound our results by giving production participants a segmentation advantage (in that they see each syllable individually on its own button), we made the sentence in the comprehension task appear on screen one syllable at a time, with a new syllable appearing every 500 ms. Once the full sentence was visible, participants could click on one of the two scenes. They received feedback on their response which stayed on-screen until they clicked 'next' to move to the next trial. Each participant did 18 comprehension trials, one per familiar scene, shown in a random order.

2.2.3 Testing

After the practice phase, all participants were asked to judge a number of sentences, some familiar and some novel. In each trial, participants saw a scene and a sentence, along with the prompt "Could someone who speaks this language describe this scene using the sentence below?". We used the `f` and `j` keys for "yes" and "no", with the mapping randomly determined for each participant (but kept the same for each trial). Participants received no feedback during this phase: pressing either `f` or `j` immediately moved them on to the next trial.

The testing phase contained four kinds of trials. First, there were GRAMMATICAL trials: nine of the familiar scenes and sentences from training, randomly sampled. If participants learned the language, they should always accept these sentences. Second, there were UNGRAMMATICAL trials: the other nine familiar scenes from training, but with sentences rearranged into a different word order (SVO, rather than the SOV participants were trained on). If participants learned the word order rule in the language, we reasoned that they should always reject these sentences.

We preregistered a particular exclusion criterion for these sentences which allowed participants to make up to and including four mistakes across these 18 GRAMMATICAL and UNGRAMMATICAL trials. In other words, the minimum accuracy permitted was 77.7%. However, it is worth noting that by excluding participants who accepted a different word order, we might be rejecting exactly those participants who had adopted a case marking analysis, since case marking languages generally permit freer word order (Fedzechkina et al., 2011; Lupyan and Dale, 2010). In an exploratory analysis reported in Appendix A, we removed the ungrammaticality criterion and re-ran the analysis we describe below, this time including participants who accepted any number of "ungrammatical" sentences. The overall pattern of results remains the same regardless of whether we use this criterion. This suggests that participants who accept the "ungrammatical" SVO sentences do so not because they have learned a free word order along with a case marking rule, but because they haven't learned the language reliably.

The final two trial types in the testing phase provide the critical data for our research question. In both trial types, the scenes contained familiar characters, but their thematic roles are reversed from the ones participants saw them in during training. Reversing the thematic roles causes the segmented analysis to yield a different sentence than the unsegmented analysis.

To illustrate: if a participant learned that the sentence *fuvu zijo gix* goes along with the fairy pushing the doctor, then in the testing phase, they would encounter two trials with a scene of the doctor pushing the fairy. In the SEGMENTED trial, they would see the doctor pushing the fairy along with the sentence in (3), which was formed by swapping just the CV stems. In the UNSEGMENTED trial, they would see this same scene along with the sentence in (4), which was formed by swapping the entire nouns. Participants were asked to judge novel sentences formed according to these two rules for all 18 reversed-role scenes.

(3) zi-vu fu-jo gix
doctor-NOM fairy-ACC push

(4) zijo fuvu gix
doctor fairy push

All in all, the testing phase contained 54 trials (9 GRAMMATICAL + 9 UNGRAMMATICAL + 18 SEGMENTED + 18 UNSEGMENTED). The order of these trials was randomised for each participant.

2.2.4 Held-out character naming

The final phase of the experiment involved a one-shot trial in which participants saw a scene with one familiar character, one held-out character that had not been previously seen in the experiment, and a familiar action happening between them. These elements were all randomly chosen. The familiar character always appeared in the same thematic role from training, so the label for that character was also familiar. The held-out character assumed the other role.

Along with the scene, participants saw a sentence with a gap where the word for the new character would be. They were asked "What seems like the most plausible word for the new character in this scene?". Two alternatives were provided, formed by combining a random held-out stem with *-vu* and with *-jo*. For example, if the scene was the fairy (familiar noun) pushing the wizard (unfamiliar noun), and the sentence was *fuvu* ____ *gix*, participants would be asked to choose between *kivu* and *kijo* as the label for the wizard.

2.3 Participants and exclusions

We used Prolific to recruit 183 adults resident in the UK who self-reported that their first language was English and that they had no known language disorders. They all gave informed consent to participate in the experiment.

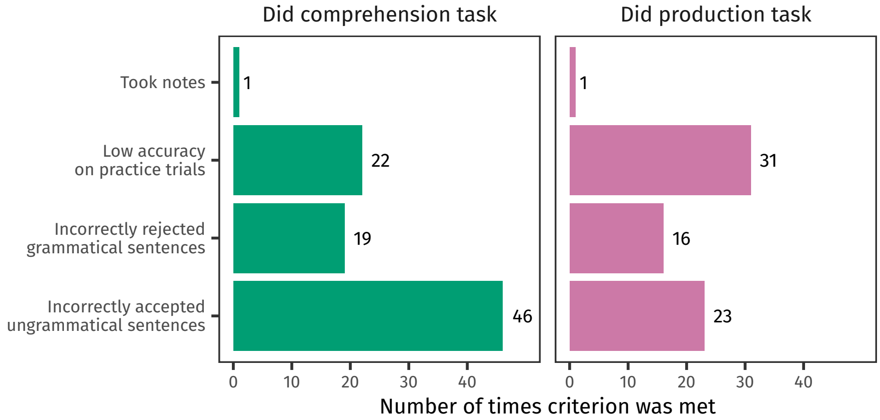

Figure 2: How many times each preregistered exclusion criterion was met in Experiment 1 (participants caught by more than one criterion contribute to each criterion's count). On the whole, exclusion criteria were met more often in the COMPREHENSION condition than in the PRODUCTION condition.

The experiment took around 20 minutes to complete (median time = 17:38), and participants were paid £3.50 (above UK National Minimum Wage at the time of running the experiment). Participants were randomly assigned to either the COMPREHENSION condition (100 people) or the PRODUCTION condition (83 people). We excluded 103 participants for the following preregistered reasons: self-reporting the use of written notes in an exit questionnaire contrary to instructions (2); low accuracy (< 77.7%, i.e., 14/18) on practice trials (16), testing trials (49) or both (36).

Figure 2 illustrates how many times each exclusion criterion was met in each condition. This plot does not reflect how the criteria may overlap, so participants caught by multiple criteria contribute to multiple counts; see Appendix B for a full breakdown of how many participants were caught by each combination of criteria.

We had to exclude many more participants who had been originally recruited into the COMPREHENSION condition, and fewer who were recruited into the PRODUCTION condition. This asymmetry indicates at least anecdotally that the production task does seem to have helped participants learn the sentences that they were exposed to—in line with previous evidence that production is good for learning.

After exclusions, we were left with analysable data from 40 participants in each condition. (Appendix C contains the same analysis that we report below run on all 183 participants.) The remaining participants' accuracy on the grammatical and ungrammatical sentences was all close to ceiling (naturally, since these are the participants who were not excluded for low accuracy), and there were no substantial differences between conditions. For the COMPREHENSION group, grammatical sentences were correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 98% of the time. And for the PRODUCTION group, grammatical sentences were also correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 97% of the time.

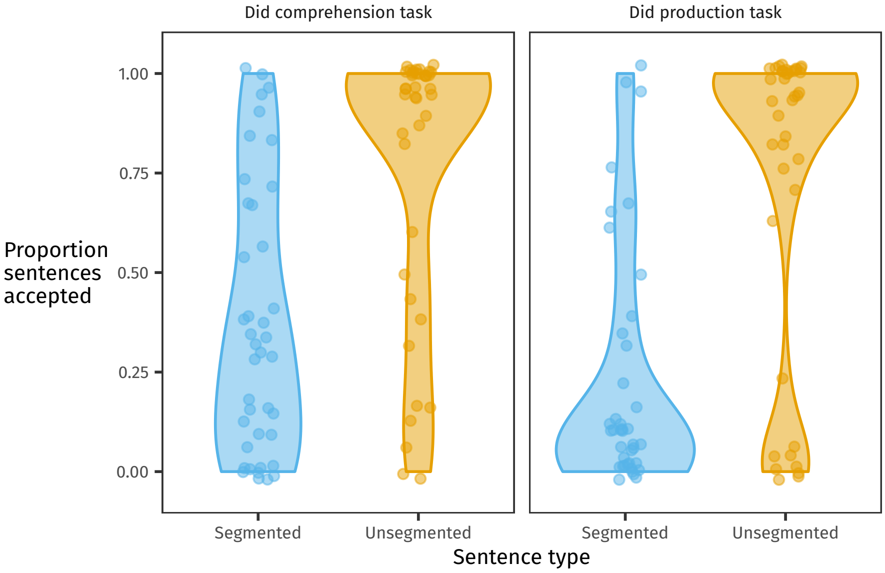

Figure 3: Participants in both the COMPREHENSION and PRODUCTION conditions of Experiment 1 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis. Each dot represents one participant's proportion of accepted sentences of each type.

2.4 Results

2.4.1 Judgement

Participants in both conditions tended to accept novel sentences formed using the unsegmented analysis, and they were more ambivalent about novel sentences formed using the segmented analysis. Figure 3 shows the proportion of each kind of novel sentence that participants accepted.

Following our preregistered analysis plan, we used brms (Bürkner, 2017) in R (R Core Team, 2024) to fit a Bayesian linear model with a Bernoulli likelihood to this data. This model predicts sentence acceptance as a function of condition (COMPREHENSION versus PRODUCTION), sentence type (SEGMENTED versus UNSEGMENTED), and their interaction. The group-level effects in the model included varying intercepts by participant and varying slopes over sentence type by participant. We selected the model's weakly regularising priors using prior predictive checks. The model converged, as indicated by all Rhats = 1.00. Appendix D contains the full model specification.

We sum-coded condition (COMPREHENSION as −0.5, PRODUCTION as +0.5) and sentence type (SEGMENTED as −0.5, UNSEEGMENTED as +0.5). The interaction term was also scaled to ±0.5 so that we could use the same weakly regularising prior for all three predictors.

We hypothesised that, if a production task helps participants learn morphological rules, then participants in the PRODUCTION condition would be more likely to accept sentences generated by the segmented analysis than participants in the COMPREHENSION condition. We would see this in the model as an interaction between condition and sentence type.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.54	0.27	0.03	1.10
Condition	−0.81	0.51	−1.79	0.20
Sentence type	4.10	0.70	2.76	5.51
Condition:Sent. type	0.37	0.66	−0.88	1.70

Table 1: The posterior probability distributions estimated by the model for the English participants' sentence acceptance data in Experiment 1. Values are on the log-odds scale.

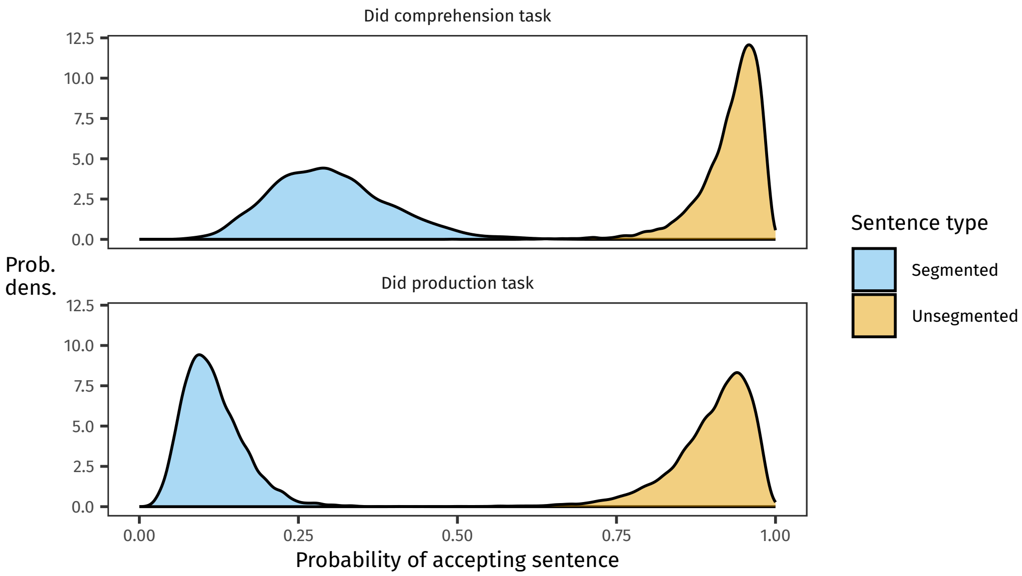

Figure 4: Conditional posterior probability distributions of the probability of accepting a sentence in Experiment 1. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

The model's posterior estimates for the population-level effects are summarised in Table 1. Figure 4 shows the conditional posterior probability distributions—that is, the posterior distributions over the probabilities of accepting a sentence for all combinations of condition and sentence type.

Overall, the model indicates with high certainty that participants are more likely to accept a novel sentence formed with the unsegmented analysis compared to a novel sentence formed with the segmented analysis. Concerning condition, the model's estimates indicate uncertainty about a difference in sentence acceptance probabilities between the PRODUCTION condition and the COMPREHENSION condition, as well as uncertainty about the interaction that our hypothesis targeted. Our prediction that participants who did the production task would be more likely to accept the novel segmented sentences was not borne out, and in fact, the results tend slightly in the opposite direction.

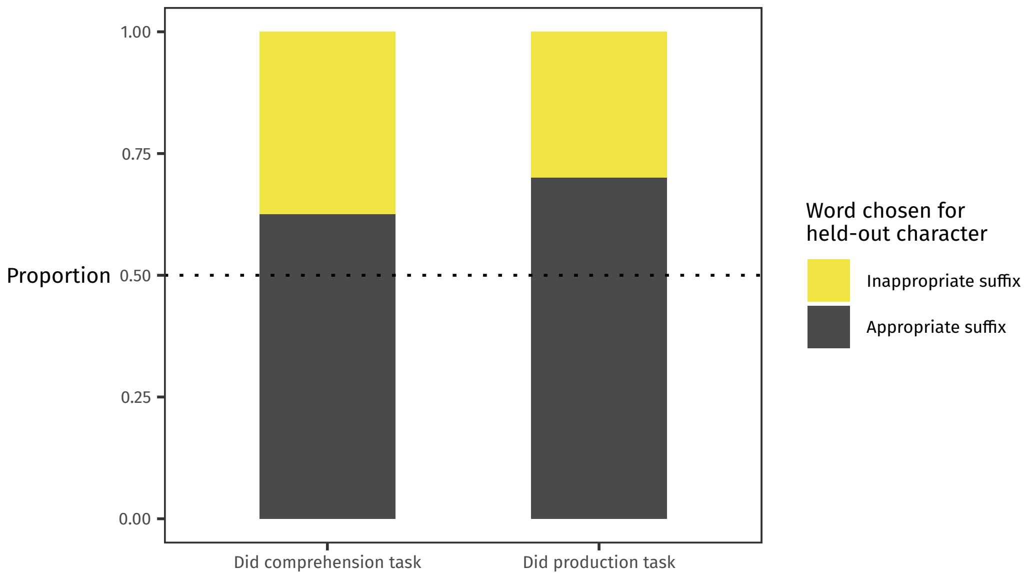

Figure 5: In the held-out character naming task of Experiment 1, more than half of participants in each condition selected the word with the appropriate suffix. Slightly more participants in the PRODUCTION condition selected the appropriate suffix.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.68	0.24	0.23	1.16
Condition	0.33	0.47	−0.57	1.22

Table 2: The posterior probability distributions estimated by the model for the English participants' held-out character naming data in Experiment 1. Values are on the log-odds scale.

2.4.2 Held-out character naming

Figure 5 shows the proportion of participants who chose a label for the held-out character that contained the appropriate suffix. Even if they didn't arrive at a fully-fledged case marking analysis, more than half of the participants in each condition seem to have noticed that each noun reliably ends in a particular syllable.

We preregistered the analysis of this data as exploratory. To see whether participants in the PRODUCTION condition showed greater awareness of these morphological patterns (even if they did not analyse them as case markers *per se*), we fit a Bayesian linear model with a Bernoulli likelihood to this data, predicting appropriate suffix choice as a function of condition (COMPREHENSION coded as −0.5, PRODUCTION as +0.5). Every participant gave only one data point, so no group-level effects were needed. We used the same weakly regularising priors as in other Bernoulli models reported in this paper. The model converged, as indicated by all Rhats = 1.00.

Table 2 summarises the posterior distributions of the population-level effects estimated by this model, and Figure 6 shows the conditional posterior probability distributions over the probabilities of selecting the appropriate suffix.

The model indicates that participants in both groups chose the label containing the

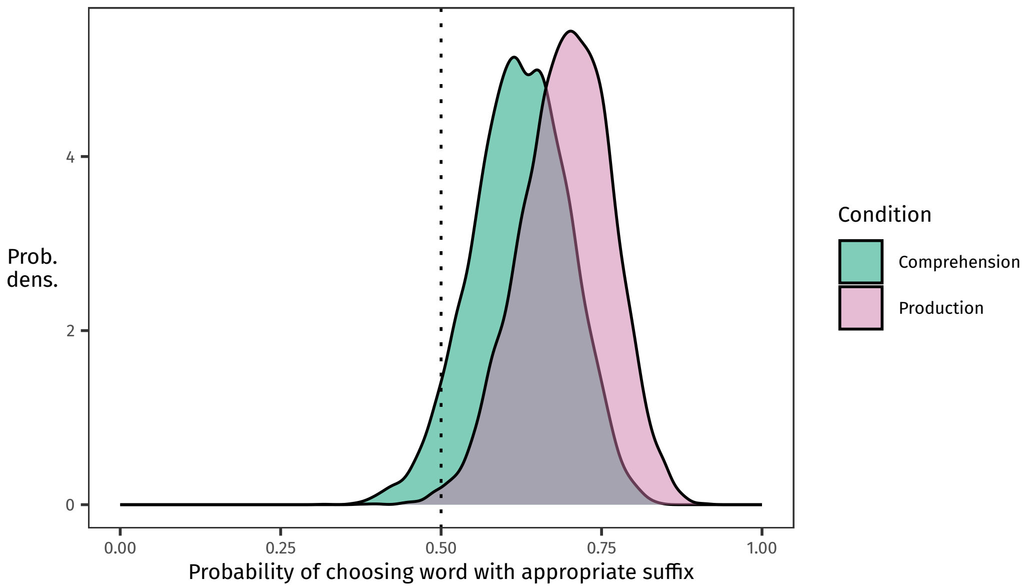

Figure 6: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 1. The overlap of these posteriors suggests uncertainty about whether and how much the groups might differ.

appropriate suffix for the missing word with a probability slightly greater than chance. Although being in the PRODUCTION condition is associated with a slightly higher probability of choosing the appropriate label, there is a great deal of overlap between conditions and thus a great deal of uncertainty about whether participants in either condition are more likely to select the appropriate label.

2.5 Interim discussion

Experiment 1 has shown that participants in both groups overwhelmingly preferred novel sentences formed using the unsegmented analysis over sentences formed using the segmented analysis. This preference was unaffected by whether participants completed a production or comprehension practice task, counter to our hypothesis.

Interestingly, the preference for the unsegmented analysis was resounding, even though the held-out character naming task indicated that many participants were aware of a morphological pattern in the language—namely that the first noun always ends in a particular syllable (the nominative marker) while the second noun always ends in another one (the accusative marker).

One straightforward explanation for this result is that the L1 English participants in this experiment might not have arrived at a case marking analysis because case is not morphologically marked outside of the pronominal system in English. In other words, English uses word order alone to indicate grammatical roles, and thus our participants may have been particularly unlikely to look beyond word order to notice that the case-marking suffixes also indicated these roles. We collected data about further languages that participants know or understand, and, in an exploratory analysis, compared the performance of participants who do know a case-marking language (15 people) to those who do not (65). The pattern of results remains the same; see Appendix E for details.

Nonetheless, it is possible that a population whose L1 includes more widespread use of case would be more likely to access the case marking analysis. We therefore ran a follow-up experiment with L1 speakers of German, a language with a productive case marking system featuring (among other cases) nominative and accusative differentially marked on nominal dependents like determiners.

3 Experiment 2

3.1 Materials

We used largely the same materials as in Experiment 1, described above in Section 2.1. Only a handful of changes were made for German-speaking participants.

First, we removed any forms from the language that resembled German words: *zug* is like German *Zug* 'train', *kex* might be read as *Keks* 'cookie', and so on.

Second, to ensure that the full set of stimuli was grammatically equivalent in German, we removed all images containing the pirate character. The German word *Pirat* is a so-called "strong masculine" noun: a noun that itself inflects for case, in addition to the usual inflection on the determiner (cf. nominative *der Pirat* 'the pirate', accusative *den Piraten*). All other characters correspond to German nouns that are grammatically "weak", that is, the nouns don't inflect for case.

Third, we changed the default word order from SOV to VSO, because SOV is the basic word order of German (Haftka, 1996; Haider, 2020). This means that the Experiment 1 sentence *fuvu zijo gix* would become *gix fuvu zijo* in Experiment 2. The "ungrammatical" word order in the judgement phase remained SVO, akin to German's V2 (though we did not use rejection of SVO sentences as a criterion for excluding participants in Experiment 2; we will discuss this further in Section 3.3).

3.2 Procedure

Experiment 2 followed the same procedure as Experiment 1 (see Section 2.2), with one modification. For English participants, we had randomly mapped the keys `f` and `j` to 'yes' and 'no'. Since German *ja* 'yes' begins with J, we instead used `p` and `q` as the decision keys for the sentence judgement task.

3.3 Participants and exclusions

We used Prolific to recruit 135 participants who self-reported that their first language was German and that they had no known language disorders. They all gave informed consent to participate in the experiment.

The experiment took around 20 minutes to complete (median time = 17:39), and participants were paid £3.85 (approx. €4.50), above UK National Minimum Wage at the time of running the experiment. As in Experiment 1, participants were randomly assigned to either the COMPREHENSION condition (68 people) or the PRODUCTION condition (67 people). We excluded 43 participants for the following preregistered reasons: low accuracy on practice trials (17), GRAMMATICAL testing trials (12), or both (14). Figure 7 illustrates

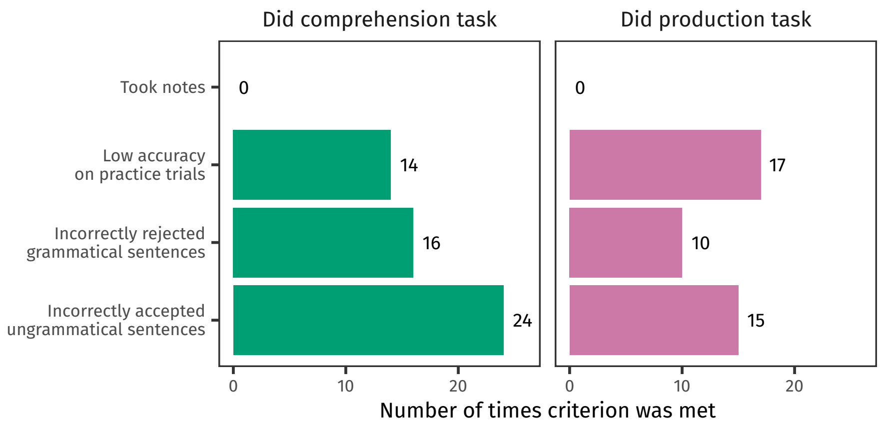

Figure 7: How many times each preregistered exclusion criterion was met in Experiment 2 (participants caught by more than one criterion contribute to each criterion's count). Exclusions are more balanced between conditions in Experiment 2 compared to Experiment 1, though still, more participants in the COMPREHENSION group compared to the PRODUCTION group incorrectly rejected sentences that were grammatical. (The ungrammatical sentences criterion is included here only for completeness; in Experiment 2 it was not used to exclude participants.)

how many times each exclusion criterion was met in each condition (note that this plot does not reflect how criteria may overlap, so participants caught by multiple criteria contribute to multiple counts).

This figure includes German participants' performance on the so-called "ungrammatical" sentences, the ones with word order that differs frrom training, though we did not use this criterion to exclude participants from the analysis. We ignored this criterion for German speakers because German permits a relatively free word order, so participants may not have had the expectation that word order should be fixed, particularly if they accessed the segmented (case marking) analysis. Recall that removing this criterion for the English-speaking participants in Experiment 1 did not affect the pattern of results (Appendix A).

After exclusions, we were left with data from 46 participants in each condition. The remaining participants' accuracy on the grammatical and ungrammatical sentences was fairly high, with no substantial differences between conditions. For the COMPREHENSION group, grammatical sentences were correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 78% of the time. And for the PRODUCTION group, grammatical sentences were also correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 82% of the time.

3.4 Results

Overall, the results from the German participants in Experiment 2 are similar to the results from the English participants in Experiment 1.

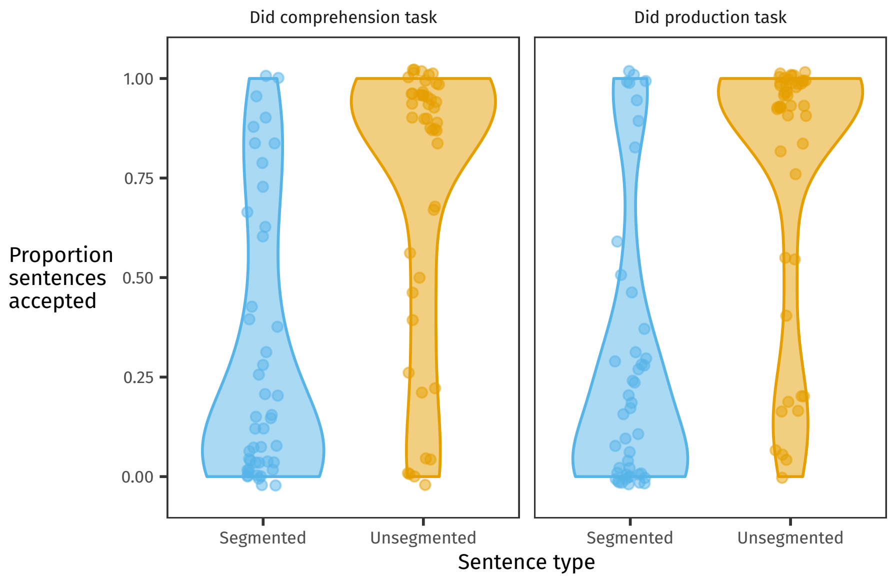

Figure 8: In Experiment 2, participants in both the COMPREHENSION and PRODUCTION conditions again accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.17	0.21	−0.24	0.59
Condition	0.33	0.40	−0.45	1.13
Sentence type	3.84	0.59	2.68	5.01
Condition:Sent. type	0.52	0.58	−0.61	1.68

Table 3: The posterior probability distributions estimated by the model for the German participants' sentence acceptance data in Experiment 2. Values are on the log-odds scale.

3.4.1 Judgement

Like the English-speaking participants, the German participants showed a clear preference for the unsegmented analysis (see Figure 8), even though German has productive morphological case marking. We fit the same Bayesian linear model as described above in Section 2.2.3 to the data from the German participants. The posterior distributions for the population-level effects estimated by the model are given in Table 3, and the conditional posterior probability distributions are shown in Figure 9. Again, the interaction that would support our hypothesis about a production task enabling participants to learn the segmented analysis was not borne out.

3.4.2 Held-out character naming

About three-quarters of German participants appear to have noticed that one noun always ends in -*vu* and the other always ends in -*jo*; see Figure 10.

We fit the same model as described in Section 2.2.4 to this data. Table 4 summarises the posterior distributions of the population-level effects, and Figure 11 shows the conditional posterior probability distributions over the probabilities of selecting the appro-

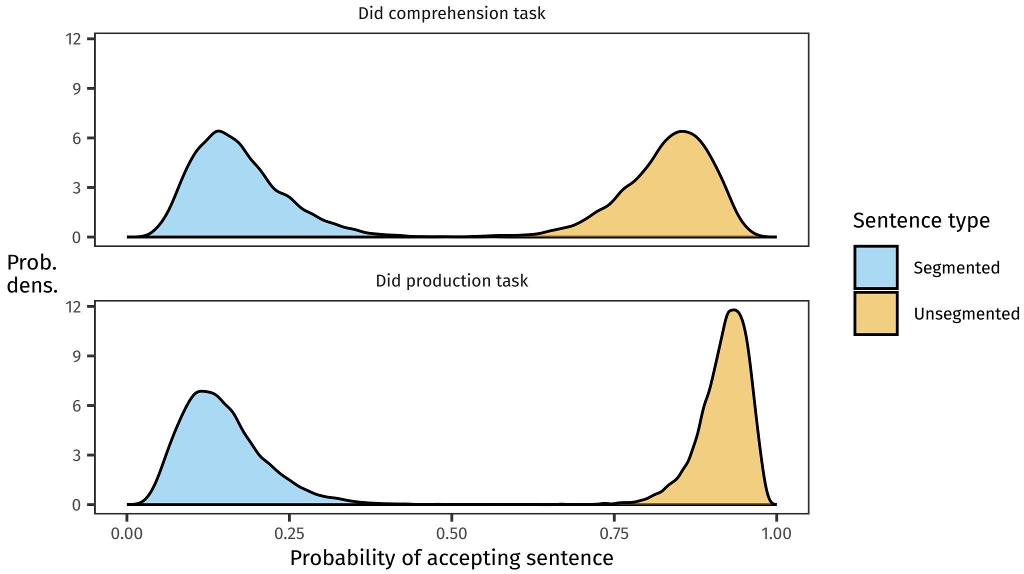

Figure 9: Conditional posterior probability distributions of the probability of accepting a sentence for the participants in Experiment 2. As in Experiment 1, UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

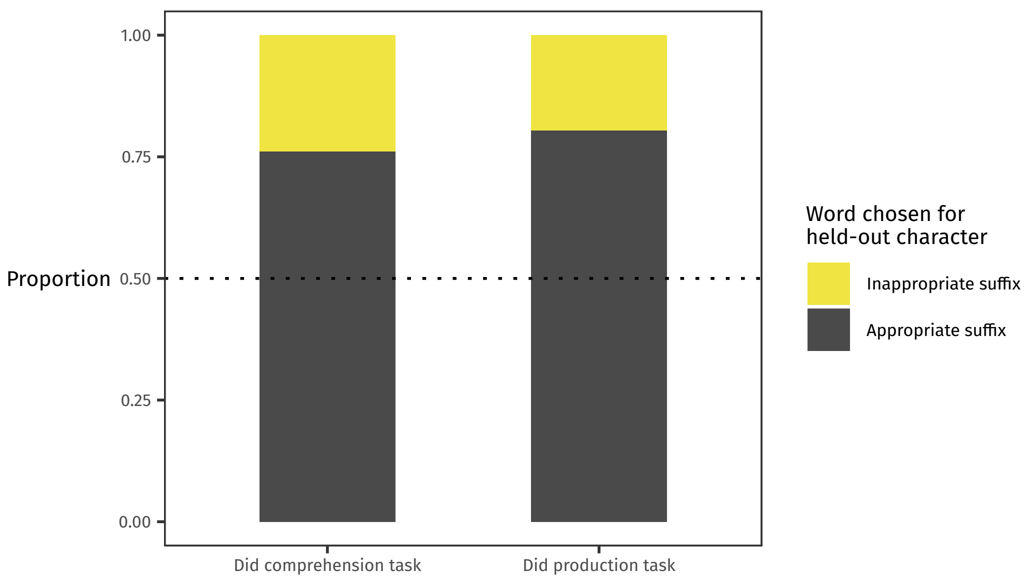

Figure 10: In the held-out character naming task of Experiment 2, around three-quarters of German participants selected the form in which the word ended in the appropriate suffix; the proportion of appropriate choices is slightly higher for the production group.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	1.28	0.25	0.80	1.79
Condition	0.24	0.50	−0.73	1.22

Table 4: The posterior probability distributions estimated by the model for the German participants' held-out character naming data in Experiment 2. Values are on the log-odds scale.

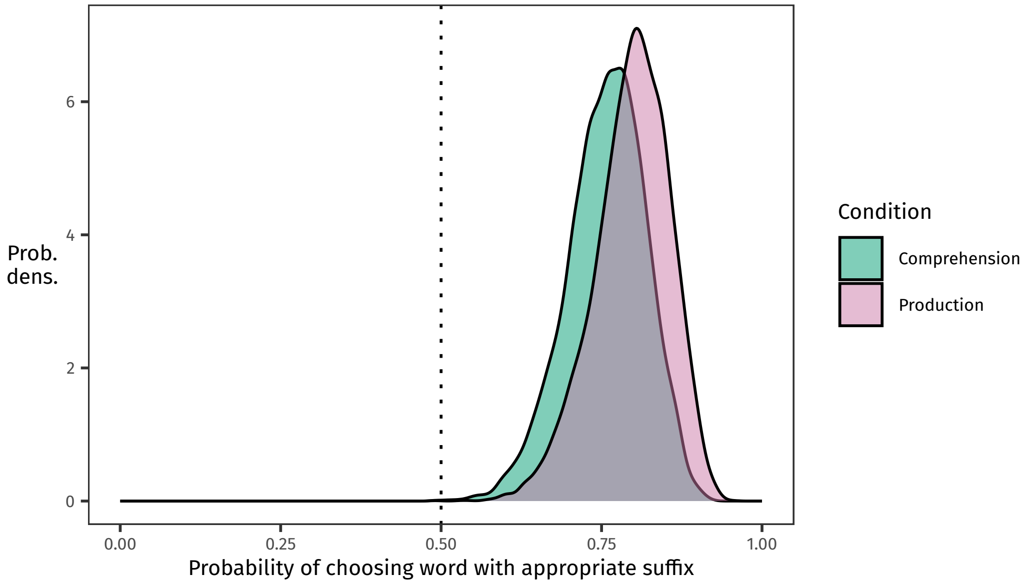

Figure 11: Conditional posterior probability distributions over the probability of German participants selecting a word that contains the appropriate suffix in Experiment 2. These posteriors overlap, so we are not certain whether and how much the groups might differ.

priate suffix. The model suggests that, much like the English participants, the German group is likely to have labelled the held-out character following the morphological pattern, and there is no clear association between participants' choice of label for the held-out character and experimental condition.

4 General discussion

In two artificial language learning experiments, we tested whether a production task—known to improve rule learning in a number of contexts—could also draw adult learners' attention to rules that adults typically disprefer. Specifically, we focused on morphological marking of thematic roles using case suffixes.

We trained participants on a language with fixed word order in which agent nouns and patient nouns were always marked with distinct suffixes (e.g., -*vu* for agents and -*jo* for patients). However, the nouns that each participant saw only ever occurred as either agents or patients, never in both roles. Thus the suffixes could be analysed as part of the nouns themselves (an unsegmented analysis) or as productive endings, part of a wider case system (a segmented analysis).

We found that regardless of whether participants did a production or a comprehension practice task, they favoured novel sentences which were formed using an unsegmented, word-level analysis, and they tended to reject sentences formed using a segmented, case-marking analysis. In other words, when shown novel scenes in which familiar characters featured in a novel grammatical role (e.g., where the fairy, which appeared only as an agent in training, appeared as the patient), they tended to reject sentences in which the noun suffixes were adjusted to reflect these new grammatical roles. Nevertheless, most participants detected the morphological pattern that resulted from the case marking (i.e., that one noun in every sentence ended in *vu* and the other in *jo*), even if they did not necessarily develop this observation into a productive case marking grammar. Perhaps surprisingly, we found that the same pattern of results—sensitivity to the morphological patterns but failure to accept sentences formed according to the segmented analysis, regardless of practice condition—also held for participants whose first language, German, has extensive case marking.

In a sense, reanalysing an unsegmented word-order-based grammar into a case marking grammar is not a trivial task, since it means overriding the chunks that have already been learned. But it is something that learners of genuine case marking languages are likely to need to do—many nouns are more likely to occur in a particular grammatical role, e.g., humans and other animate beings are more commonly found as agents than as patients (Croft, 2003; Meir et al., 2017; Silverstein, 1976). So it is not unreasonable to expect our participants to be able to break down the chunks they have learned.

In short, then, we have found no evidence that production tasks have an advantage over comprehension tasks for helping adult learners acquire a more difficult morphological rule over a more available word-level one.

Of course, as with any comparison of comprehension and production, it is difficult to be sure that we have isolated the relevant mechanism that makes production help learners. For example, one potential criticism of our study is that our production task required participants to click on buttons to build up a sentence syllable-by-syllable, rather than to produce the sentence aloud themselves. Perhaps this was not active enough to elicit the benefits of language production that previous research describes. However, we find this explanation unlikely. As mentioned above, studies on the effect of different kinds of tests (e.g., short answer, multiple choice) have found that any kind of testing can improve learning over a passive rereading or recognition task (Kang et al., 2007; McDaniel et al., 2007; McDermott et al., 2014). These studies do suggest, however, that the degree of improvement may not be the same between all kinds of test. So perhaps the kind of production task we did only got participants part of the way, not as far as they may have come with from-scratch verbal production.

In our view, a more plausible explanation for why we failed to find an improvement with production is that participants were not required to produce the language early enough in the learning process: the critical practice phase came after an initial training phase. We designed the task in this way because we were concerned about disproportional attrition of participants in the PRODUCTION condition compared to the COMPREHENSION condition. Constructing sentences in an unfamiliar language is a much more challenging task than choosing between pictures, and we didn't want PRODUCTION participants to be discouraged (and potentially stop the study at disproportionate rates) by introducing this extra level of difficulty too early. However, introducing the different practice tasks too late also has a downside: participants may have already discovered the fixed word order rule during the training phase, and since that rule perfectly explains all the data they encountered, there was no need to search for further explanations (in classical conditioning terms, an *overshadowing* effect; Pavlov, 1927). This pattern of behaviour is characteristic of adults in non-linguistic tasks too: adults tend to identify a reliable cue and then exploit it, while children continue to explore (Liquin and Gopnik, 2022; Sumner et al., 2019). A possible prediction of this account, then, is that children might be more likely than adults to accept the case marking analysis.

The late start of the production/comprehension tasks could also be part of why our results differ from those of Hopman and MacDonald (2018), who observe that a production task leads to slightly better learning of morphological rules than word order rules. In their design, passive exposure trials were interleaved with blocks of active production trials. And their experiment seems to have been conducted in person, a factor likely to prevent participants from withdrawing from the experiment early, compared to experiments run online.

4.1 Starting big, but noticing small

Our results, and previous findings showing that adults struggle to learn morphological rules, align with the observation that as language is learned, linguistic information tends to be stored first as holistic chunks (Christiansen and Chater, 2016). The morphosyntactic rules that might underlie parts of those chunks are only induced when learners receive sufficient evidence from the input. This idea has been referred to as "needs-only analysis" (Wray, 2002, 2006), and it is closely related to the "starting big" approach described by Havron and Arnon (2021) and Siegelman and Arnon (2015).

We do see our adult participants "starting big" by learning a rule that manipulates the larger, unsegmented units, not the smaller ones that require word segmentation. But our results from the held-out character naming task also add some nuance to this notion. Participants still notice patterns and pieces within the word-level chunks they manipulate. In other words, the chunks they learn aren't fully opaque.

This is interesting in connection with anecdotal evidence of many potentially segmentable chunks being learned holistically. For example, people might have a moment of surprise when they realise, say, that a safety pin is so called because it is safe, or that dry cleaning is a kind of cleaning which doesn't involve water. Those situations are evidently different from the suffixing pattern in our artificial language. The suffixes seem to have been salient enough for people to notice, even if the noticing doesn't cause learners to override the chunk they learned.

By noticing this pattern at all, though, learners have taken the first step toward such an analysis. How might we push them to make the leap? Evidence from many learning domains suggests that learners need more data—more unique examples of a rule in action, in varying contexts—to move beyond item-by-item learning to systematic rule induction (see Raviv et al., 2022 for a review). When the rule to be acquired is dispreferred *a priori*, either more data or conflicting data (see next section) may be required to overcome participants' strong prior preferences.

However, there is a trade-off here. More variable input does lead to better long-term generalisation, but it also hinders initial learning (Raviv et al., 2022). And in an experimental setting, these may be hard to balance. Logistics and finances limit how much training participants can receive before they need to provide useful data. It would be difficult within the current experimental design to include the large amount of variability required, while also ensuring that participants learned the language adequately.

4.2 Outlook and future directions

We see two interesting options for follow-up research based on our results and the observations outlined above. First and most obviously, the production and comprehension tasks could be interspersed throughout the training phase. The greater attrition rate that we would expect from this design could be handled either by significant over-recruitment of participants or in-person administration of the experiment.

Second, knowing now that adult learners will tend to adopt the unsegmented analysis, we might ask: how difficult would it be to pivot from that initial analysis to a segmented one in the face of new, conflicting data? The under-specified sentences shown during training could be changed partway through to become sentences unambiguously formed using a case marking rule, incompatible with the unsegmented analysis that participants would presumably have learned. Or alternatively, perhaps just one character alternates thematic roles, or one character is irregular and takes no suffix at all.

We could imagine two reasons why a production task might help participants more swiftly reanalyse their data in the face of this conflicting evidence. First, according to the noticing mechanism (Swain, 2005), production may help participants more quickly identify the morphological patterns that are now the only way to fully explain the data. Second, the production advantage has also been explained in terms of retrieval from memory, since retrieval practice is known to strengthen learning (Hopman and MacDonald, 2018; Karpicke, 2012; Karpicke and Roediger, 2008; MacDonald, 2013). Under this mechanism, production could help participants recall whatever units they had stored—likely unsegmented ones (Christiansen and Chater, 2016; Havron and Arnon, 2021; Siegelman and Arnon, 2015)—to render them more available for reanalysis.

5 Conclusion

We began this investigation where two strands of previous research intersect, one showing that that language production helps learners identify and learn rules in their language (Hopman and MacDonald, 2018; Izumi, 2002; Swain, 2005), and another showing that adults struggle to learn morphological rules and prefer word-level ones (Havron and Arnon, 2021; Jordens et al., 1989; Lupyan and Dale, 2010; Papadopoulou et al., 2011; Parodi et al., 2004). Bringing these observations together, we wanted to know whether a production task could help adult learners to identify a more difficult morphological rule over a more available word-level one.

Our results demonstrate that adults prefer to learn a word-level rule for marking thematic role over a morphological rule, even when they appear to notice morphological patterns. Contrary to our preregistered hypothesis, practising a new language with a production task does not steer learners away from this strong preference for word-level rules. This holds for speakers of both English (Experiment 1) and German (Experiment

2), indicating that even adult learners familiar with case marking tend to prefer the word-level rule over the morphological one.

Although we found no evidence that production tasks are better than comprehension tasks for helping adult learners acquire morphological rules, these two experiments nevertheless clearly illustrate that adults strongly prefer to learn rules operating over larger units, rather than smaller ones. This supports assumptions made in, e.g., Bentz and Winter (2013) and Lupyan and Dale (2010), for whom adult preference for word-level rules is central for explaining why languages with more adult learners tend to have less complex morphology.

All in all, although our hypothesis about the role of production for morphological learning was not borne out, this study has still opened several doors that we believe are worth passing through to discover more about the interplay of language production and the kinds of rules that learners learn.

References

Bentz, C. and Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, 3(1):1–27. 2, 3, 22, 26

Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., and Gillam, R. B. (2010). What you hear and what you say: Language performance in Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism*, 13(3):325–344. 1

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28. 10

Christiansen, M. H. and Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62. 20, 21

Croft, W. (2003). *Typology and Universals*. Cambridge University Press, Cambridge, 2 edition. 19

de Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source CollaborativeEcosystem of Behavioral Experiments. *Journal of Open Source Software*, 8(85):5351. 5

Donnelly, S. and Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, 92(2):609–625. 1

Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2011). Functional biases in language learning: Evidence from word order and case-marking interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33. 3, 7, 26

Haftka, B. (1996). Deutsch ist eine V/2-Sprache mit Verbendstellung und freier Wortfolge. In Lang, E. and Zifonun, G., editors, *Deutsch - Typologisch*, pages 121–141. De Gruyter. 14

Haider, H. (2020). VO-/OV-Base Ordering. In Putnam, M. T. and Page, B. R., editors, *The Cambridge Handbook of Germanic Linguistics*, pages 339–364. Cambridge University Press, 1 edition. 14

Havron, N. and Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, 48(2):244–260. 20, 21

Holmes, V. M. and Dejean De La Bâtie, B. (1999). Assignment of grammatical gender by native speakers and foreign learners of french. *Applied Psycholinguistics*, 20(4):479–506. 2

Hopman, E. W. M. (2022). *Modality Matters: Generalization in Second Language Learning after Production versus Comprehension Practice*. PhD thesis, University of Wisconsin-Madison. 2

Hopman, E. W. M. and MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, 29(6):961–971. 2, 20, 21

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, 24(4):541–577. 1, 2, 21

Jordens, P., de Bot, K., and Trapman, H. (1989). Linguistic aspects of regression in German case marking. *Studies in Second Language Acquisition*, 11:179–204. 2, 21

Kang, S. H. K., McDermott, K. B., and 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. 2, 19

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21(3):157–163. 21

Karpicke, J. D. and Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science (New York, N.Y.)*, 319(5865):966–968. 21

Keogh, A. and Lupyan, G. (in press). Who benefits from redundancy in learning noun class systems? In *Proceedings of the 15th International Conference on the Evolution of Language*. 2

Keppenne, V., Hopman, E. W. M., and Jackson, C. N. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics*, 42(4):907–936. 2

Levshina, N. (2020). Efficient trade-offs as explanations in functional linguistics: some problems and an alternative proposal. *Revista da ABRALIN*, 19(3):50–78. 3

Liquin, E. G. and Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, 218:104940. 20

Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, 5(1):e8559. 2, 3, 7, 21, 22

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4. 21

McDaniel, M. A., Anderson, J. L., Derbish, M. H., and Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19(4-5):494–513. 2, 19

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., and McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1):3–21. 2, 19

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Ben-Basat, A. L., Padden, C., and Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, 158:189–207. 19

Papadopoulou, D., Varlokosta, S., Spyropoulous, V., Kaili, H., Prokou, S., and Revithiadou, A. (2011). Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research*, 27:173–205. 2, 21

Parodi, T., Schwartz, B. D., and Clahsen, H. (2004). On the L2 acquisition of the morphosyntax of German nominals. *Linguistics*, 42:669–705. 2, 21

Pavlov, P. I. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford University Press, London. 20

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 10

Raviv, L., Lupyan, G., and Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Science*, 26(6):462–483. 20

Ribot, K. M., Hoff, E., and Burridge, A. (2018). Language use contributes to expressive language growth: Evidence from bilingual children. *Child Development*, 89(3):929–940. 1

Rogers, M. (1987). Learners Difficulties with Grammatical Gender in German as a Foreign Language*. *Applied Linguistics*, 8(1):48–74. 2

Sagarra, N. and Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition*, 35(2):261–290. 2

Siegelman, N. and Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, 85:60–75. 20, 21

Silverstein, M. (1976). Hierarchy of Features and Ergativity. In *Grammatical Categories in Australian Languages*, pages 163–232. De Gruyter. 19

Sumner, E. S., Li, A. X., Perfors, A., Hayes, B. K., Navarro, D. J., and Sarnecka, B. W. (2019). The exploration advantage: Children's instinct to explore leads them to find information that adults miss. *PsyArXiv*, page 11. 20

Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of Research in Second Language Teaching and Learning*, pages 471–483. Routledge, New York, 1 edition. 2, 21

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge. 20

Wray, A. (2006). Formulaic language. In Brown, K., editor, *Encyclopedia of Language & Linguistics*, pages 590–597. Elsevier, Oxford, 2 edition. 20

A Exploratory analysis: Removing the ungrammatical exclusion criterion

In the test phase of the experiment, we collected data that would inform two exclusion criteria: we would only keep participants who correctly accepted grammatical sentences and correctly rejected ungrammatical ones. We had defined “ungrammatical” as a word order that diverged from the one in training. Our reasoning was that participants should have learned that the language has SOV order, so they should reject the “ungrammatical” SVO order. Since SVO is also the basic word order of English, participants’ rejection of it would provide the strongest test that they had learned the word order of the artificial language.

However, if a language has case marking, it is likely to also have free word order (Bentz and Winter, 2013; Fedzechkina et al., 2011). It is therefore possible that participants who accepted sentences with a different word order had learned a case marking rule and associated that with a free word order, in which case our exclusion criterion would be removing exactly those participants who learned the segmented analysis we were targeting. This could explain why our results show such a strong preference for the unsegmented analysis.

Here, we lift this exclusion criterion and re-run the analyses described above. This criterion originally excluded 27 comprehension participants and 6 production participants; below we analyse data from 67 participants in the comprehension group and 46 in the production group.

A.1 Judgement

Figure 12 shows a similar pattern to Figure 3: a general preference for the novel sentences formed using the unsegmented analysis, and greater ambivalence toward ones formed with the segmented analysis.

We fit the same model described in Section 2.4 to this data. The pattern of results (shown in Table 5) remains the same as above. We conclude that the “ungrammatical” word order criterion did not exclude participants who learned a case marking rule and then extrapolated from it that word order was free.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.41	0.24	−0.07	0.89
Condition	−0.19	0.47	−1.12	0.72
Sentence type	3.48	0.54	2.43	4.58
Condition:Sent. type	0.88	0.56	−0.20	1.98

Table 5: Posterior distributions estimated by a model predicting sentence acceptance by condition, sentence type, and their interaction, now including data from participants originally excluded from Experiment 1 for rejecting sentences with a different word order than seen in training.

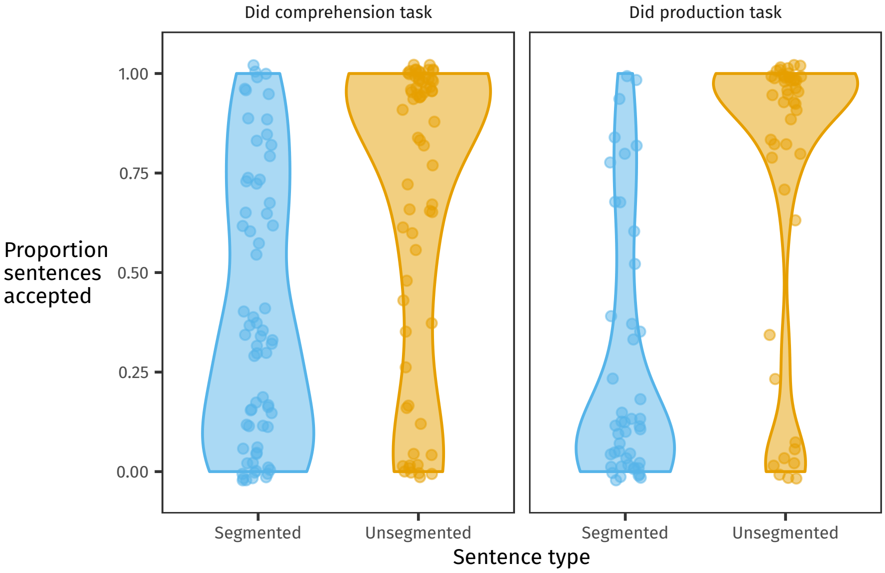

Figure 12: After lifting the ungrammaticality rejection criterion for participants in Experiment 1, the larger pool of participants show the same results: a strong preference for the unsegmented analysis over the segmented analysis, with no clear effect of task.

A.2 Held-out character naming

The results from the held-out character naming analysis also remain extremely similar to the ones reported with the original exclusion criteria, as shown in Figure 13 and Table 6.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.71	0.21	0.32	1.12
Condition	0.24	0.41	-0.55	1.06

Table 6: Posterior distributions estimated by a model predicting appropriate suffix choice by condition, now including data from participants originally excluded from Experiment 1for rejecting sentences with a different word order than seen in training.

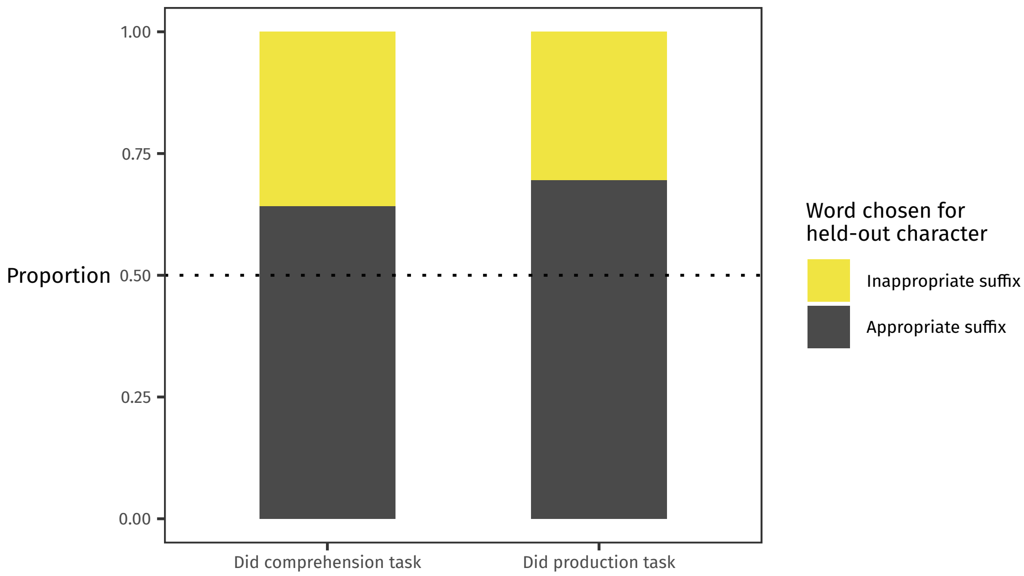

Figure 13: Proportion of Experiment 1 participants in each group, now including participants previously excluded from the analysis based on the ungrammaticality rejection criterion, who labelled the held-out character with a word containing the appropriate suffix. The same pattern holds as in the original analysis.

B Overlaps in exclusion criteria

The following table shows how many of the 183 participants recruited for Experiment 1 were caught by each combination of exclusion criteria. (Gram. = incorrectly rejected grammatical sentences; Ungram. = incorrectly accepted ungrammatical sentences; Practice = low accuracy on practice phase; Notes = self-reported taking notes.)

Gram.	Ungram.	Practice	Notes	Comprehension	Production
				40	40
			×	0	1
		×		5	12
	×			27	6
	×		×	1	0
	×	×		8	8
×				5	1
×		×		4	6
×	×			5	4
×	×	×		5	5

The following table shows how many of the 135 participants recruited for Experiment 2 were caught by each combination of exclusion criteria. (The ungrammatical sentences criterion was not used on its own to exclude participants in Experiment 2.)

Gram.	Ungram.	Practice	Notes	Comprehension	Production
				35	36
		×		4	10
	×			11	10
	×	×		2	1
×				4	2
×		×		1	4
×	×			4	2
×	×	×		7	2

C Analysis of all participants

In this appendix, we report the same analyses as in Sections 2.4 and 3.4 run on the data from all originally-recruited participants, imposing none of the preregistered criteria for exclusion.

C.1 Experiment 1

We recruited 183 participants in total for Experiment 1: 100 in the COMPREHENSION condition and 83 in the PRODUCTION condition.

C.1.1 Judgement

Figure 14 visualises the proportion of times each participant accepted each type of sentence at test. The same model described above was fit to this data; its posterior estimates are summarised in Table 7, and the conditional posterior distributions over the probability of accepting a sentence are shown in Figure 15.

Overall, we see a similar pattern to the original analysis: participants in both the COMPREHENSION and the PRODUCTION condition accept the unsegmented sentences more than the segmented sentences.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.27	0.16	−0.04	0.58
Condition	0.09	0.31	−0.54	0.69
Sentence type	2.07	0.32	1.45	2.71
Condition:Sent. type	0.23	0.32	−0.40	0.86

Table 7: The posterior probability distributions estimated by the model for the sentence acceptance data from all 183 participants recruited for Experiment 1. Values are on the log-odds scale.

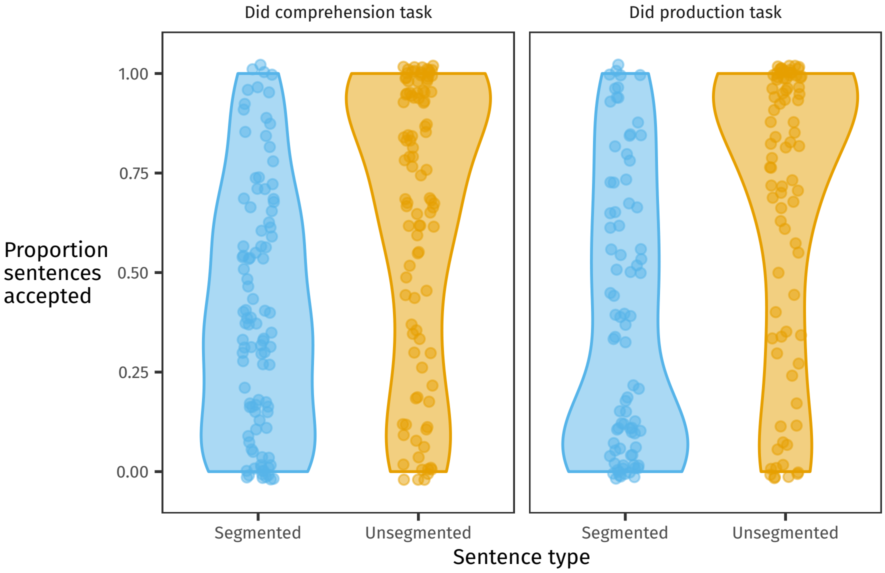

Figure 14: All 183 participants recruited for Experiment 1 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis, regardless of task. Each dot represents one participant's proportion of accepted sentences of each type.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.46	0.15	0.17	0.76
Condition	0.11	0.31	−0.48	0.72

Table 8: The posterior probability distributions estimated by the model for all 183 participants' held-out character naming data in Experiment 1. Values are on the log-odds scale.

C.1.2 Held-out character naming

Figure 16 illustrates the proportion of participants in each condition who named the held-out character using the appropriate suffix—the one that doesn't appear elsewhere in the sentence. As in the original analysis, more than half of the participants in both groups chose the word containing the appropriate, and the model estimates that both groups have very similar probabilities of selecting the appropriate suffix (see the posterior summaries in Table 8 and the conditional posterior distributions in Figure 17).

C.2 Experiment 2

We recruited 135 participants in total for Experiment 2: 68 in the COMPREHENSION condition and 67 in the PRODUCTION condition.

C.2.1 Judgement

In Figure 18, we show the proportion of times each participant accepted each type of sentence at test. We see the same pattern as in the original Experiment 2 data and in

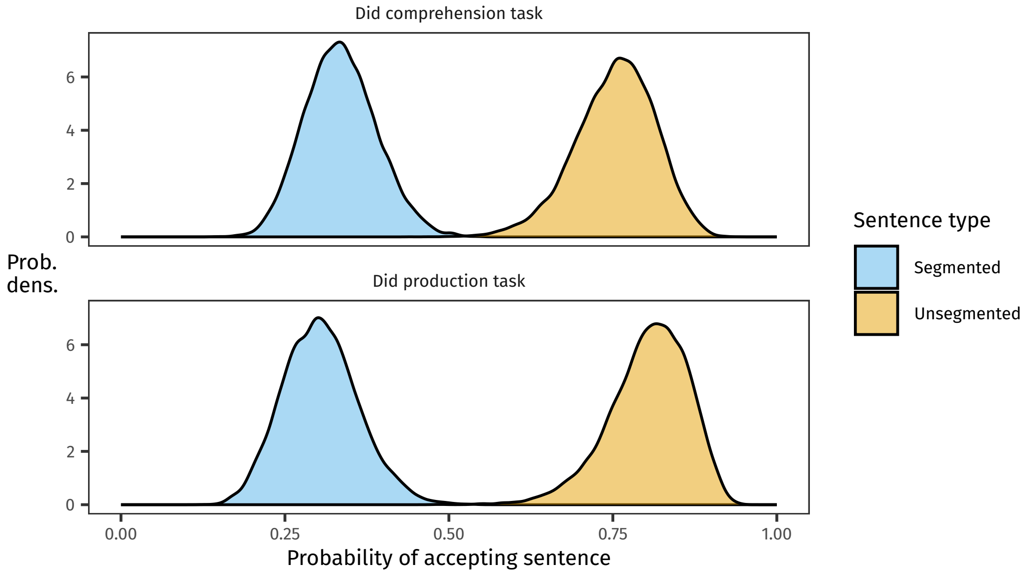

Figure 15: Conditional posterior probability distributions of the probability that all 183 participants recruited for Experiment 1 would accept a sentence. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

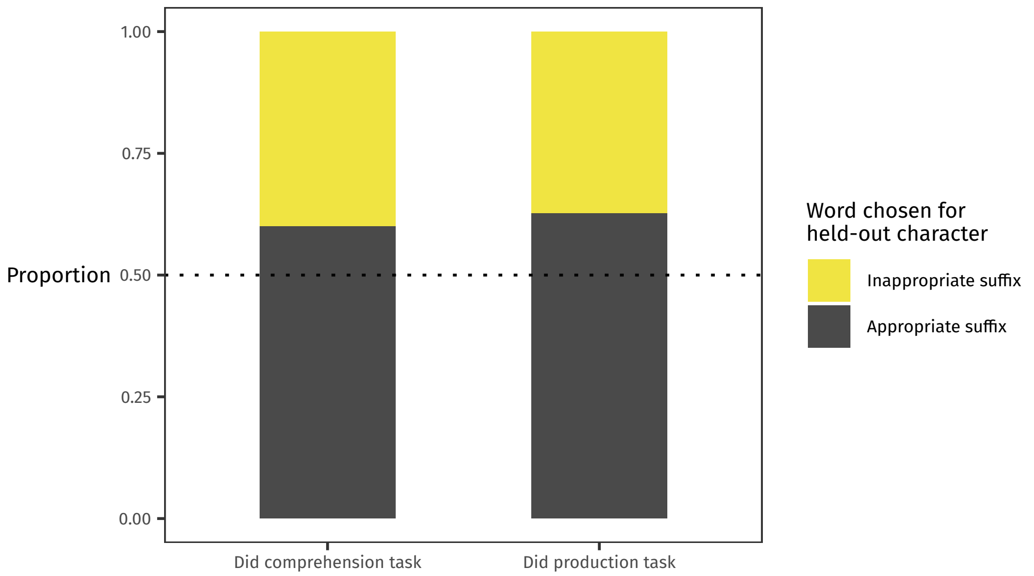

Figure 16: In the held-out character naming task of Experiment 1, more than half of all 183 participants selected the word with the appropriate suffix.

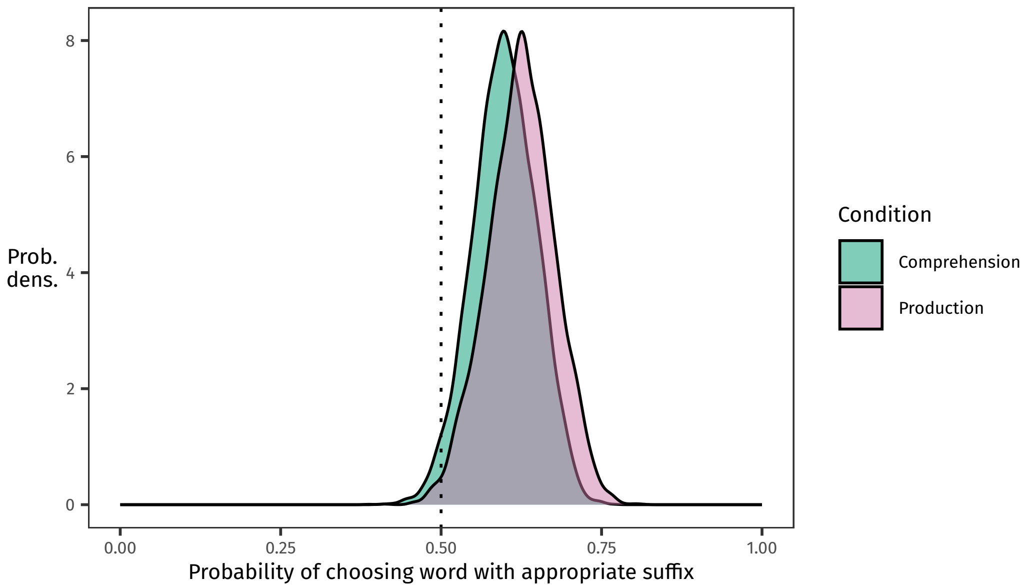

Figure 17: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 1, shown for all 183 originally recruited participants.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.15	0.14	-0.13	0.43
Condition	0.32	0.27	-0.21	0.87
Sentence type	2.51	0.41	1.72	3.33
Condition:Sent. type	0.19	0.41	-0.61	0.98

Table 9: The posterior probability distributions estimated by the model for the sentence acceptance data from all 135 participants recruited for Experiment 2. Values are on the log-odds scale.

the data of all 183 participants from Experiment 1: participants prefer the unsegmented sentences over the segmented ones, regardless of task. Table 9 summarises the posterior distributions estimated by the same model as above, and Figure 19 shows the conditional posterior distributions.

C.2.2 Held-out character naming

Figure 20 shows that, like the original analysis, around three-quarters of participants in each condition named the held-out character using the appropriate suffix. Table 10 summarises the posteriors estimated by the same model described above, and Figure 21 shows the conditional posterior distributions.

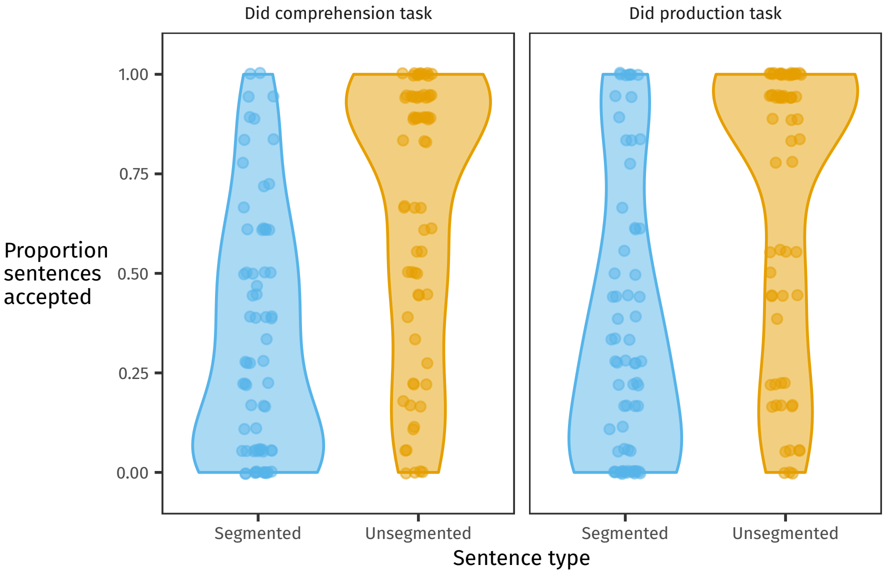

Figure 18: All 135 participants recruited for Experiment 2 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis, regardless of task. Each dot represents one participant's proportion of accepted sentences of each type.

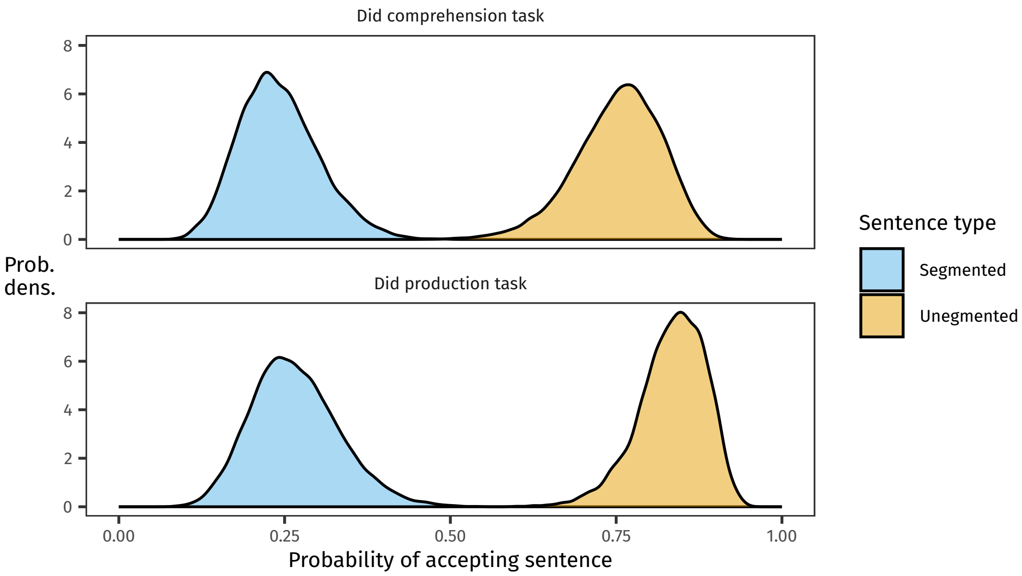

Figure 19: Conditional posterior probability distributions of the probability that all 135 participants recruited for Experiment 2 would accept a sentence. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

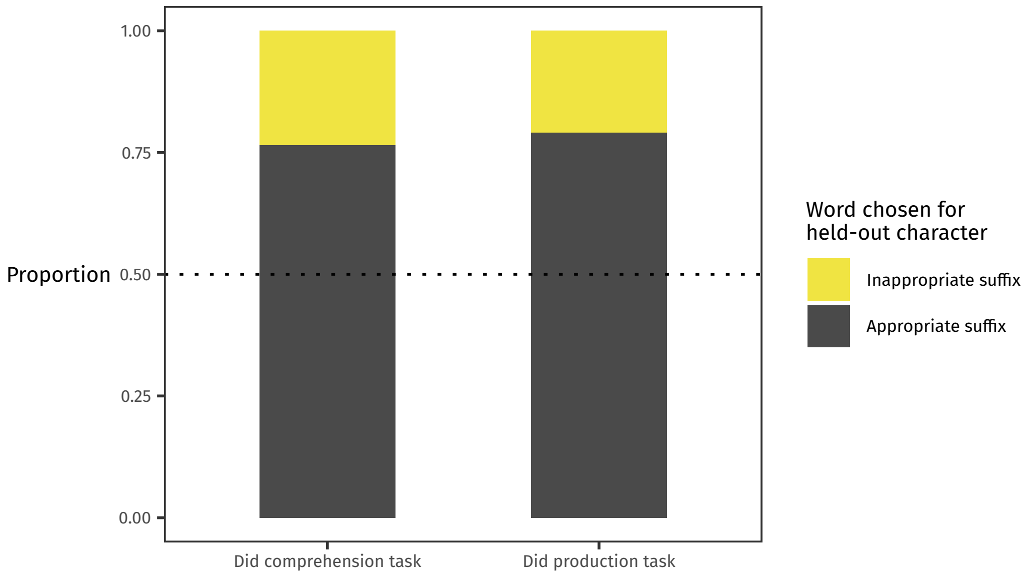

Figure 20: In the held-out character naming task of Experiment 2, at least three-quarters of all 135 participants selected the word with the appropriate suffix.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	1.26	0.21	0.86	1.67
Condition	0.14	0.41	−0.65	0.95

Table 10: The posterior probability distributions estimated by the model for all 135 participants' held-out character naming data in Experiment 2. Values are on the log-odds scale.

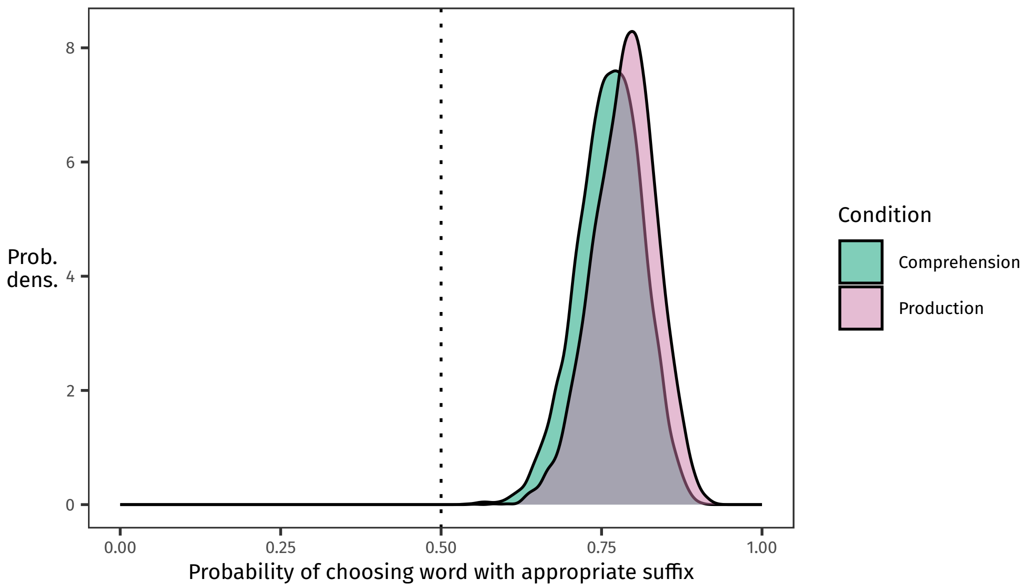

Figure 21: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 2, shown for all 135 originally recruited participants.

D Bayesian model specifications

D.1 Judgement

```
brm(
  sentence_accepted ~ sent + cond + sentcond + (sent | ppt_id),
  family = bernoulli(),
  prior = c(
    prior(normal(0, 1.5), class = Intercept),
    prior(normal(0, 2), class = b),
    prior(normal(0, 5), class = sd, coef = Intercept, group = ppt_id),
    prior(normal(0, 5), class = sd, coef = sent, group = ppt_id),
    prior(lkj(2), class = cor, group = ppt_id)
  )
)
```

D.2 Held-out character naming

```
brm(
  match ~ cond,
  family = bernoulli(),
  prior = c(
    prior(normal(0, 1.5), class = Intercept),
    prior(normal(0, 2), class = b)
  )
)
```

E Exploratory analysis: Participants who know case marking languages

In the post-experiment debrief questionnaire, we asked participants if they knew or understood any other languages beyond English. If they self-reported knowing a case marking language, we placed them into a separate group from the participants who did not. Fifteen participants out of 80 reported that they know or understand the following case marking languages: Arabic, Czech, German, Latin, Polish, Romanian, Slav, Somali, Tunisian, Turkish, and Urdu.

E.1 Judgement

Figure 22 visualises the proportion of sentence acceptance judgements for each participant, split by condition and further by whether each participant knows a case marking language.

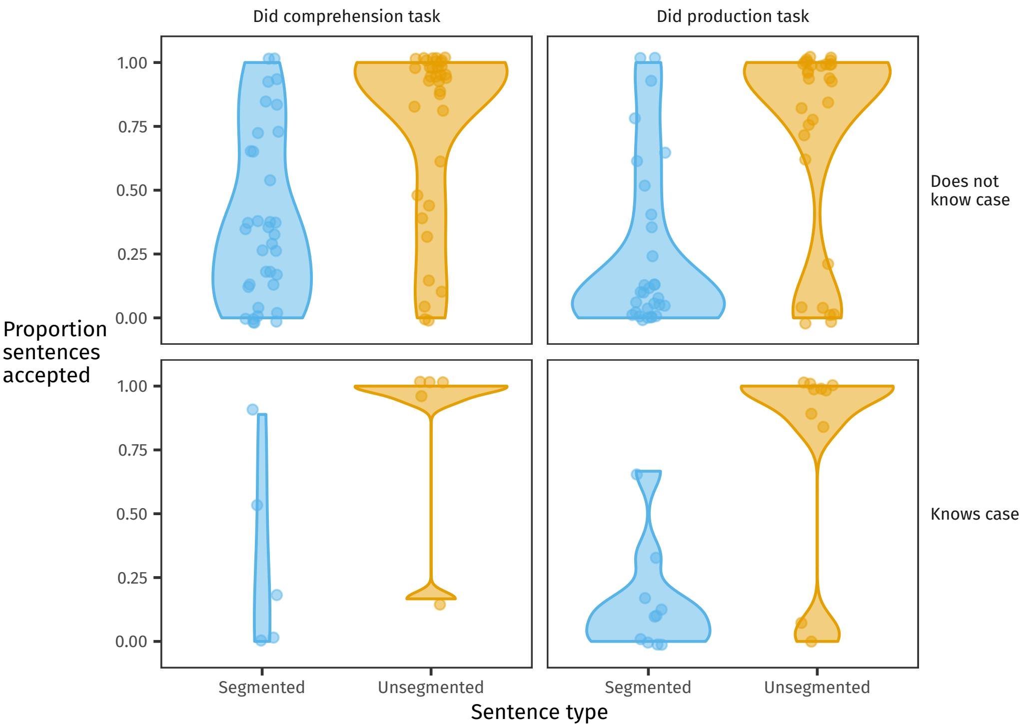

Figure 22: Participants in Experiment 1 who self-reported knowing a case marking language show a similar pattern of sentence acceptance to participants who do not know a language with case marking.

We fit the same Bayesian model as described in Section 2.4 to this data, adding in an additional sum-coded predictor for knowledge of case (−0.5 when the participant does not know a case marking language, +0.5 when they do), and all two- and three-way interactions with the predictors sentence type and condition (scaled to ±0.5). Table 11 summarises the posterior distributions of the population-level effects estimated by the model. In short, the previously-estimated effects remain qualitatively the same, and

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.53	0.35	−0.15	1.25
Condition	−0.76	0.66	−2.05	0.52
Sentence type	4.36	0.86	2.68	6.03
Case	−0.03	0.69	−1.36	1.37
Condition:Sent. type	0.52	0.84	−1.11	2.18
Condition:Case	0.09	0.67	−1.21	1.42
Sent. type:Case	0.44	0.83	−1.21	2.04
Cond.:Sent. type:Case	0.28	0.85	−1.38	1.93

Table 11: Posterior distributions estimated by a model predicting sentence acceptance by condition, sentence type, and knowledge of a case marking language, and all interactions between them.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.56	0.32	−0.06	1.19
Condition	0.10	0.62	−1.11	1.27
Case	−0.59	0.62	−1.82	0.65
Condition:Case	−0.56	0.62	−1.78	0.64

Table 12: Posterior distributions estimated by a model predicting appropriate suffix choice by condition, knowledge of a case marking language, and their interaction.

the model indicates great uncertainty about any association between prior knowledge of case marking languages and acceptance of sentences formed using the segmented analysis.

E.2 Held-out character naming

Figure 23 illustrates that the 15 participants who know a case marking language select the word with the appropriate suffix less often than the larger group of 65 participants who do not know case. However, we fit a model estimating appropriate suffix choice as a function of condition, knowledge of case, and their interaction (scaled to ±0.5), and the posterior distribution estimates in Table 12 indicate that we cannot be certain about any differences between participant groups.

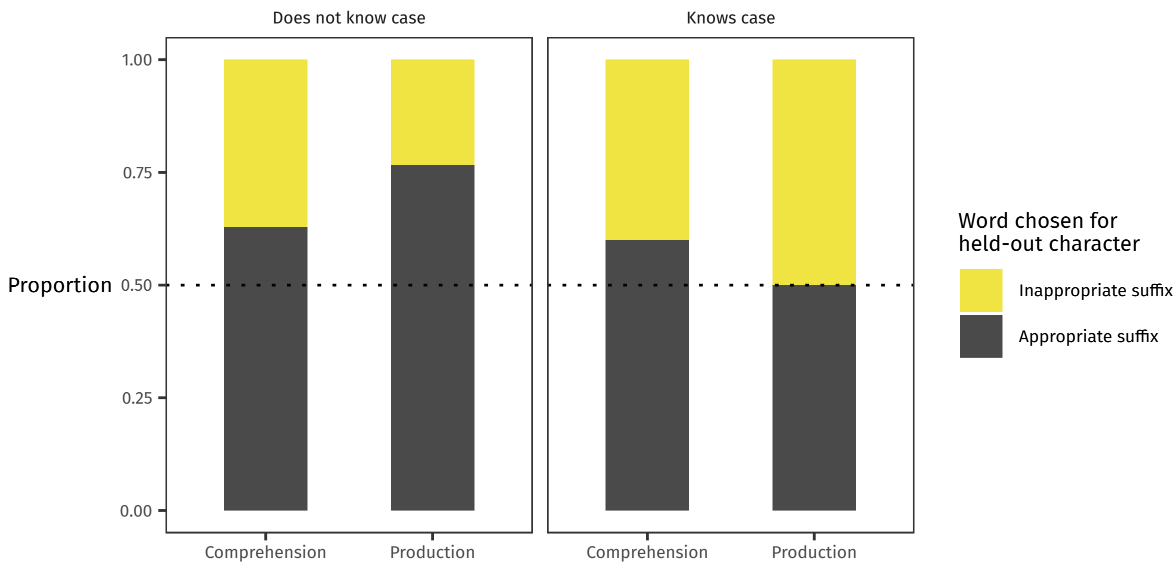

Figure 23: The 15 participants in Experiment 1 who know a case marking language give overall less appropriate responses to the held-out character naming task, with production participants selecting the appropriate suffix less than participants in the comprehension group.