COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13107

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13107

Category Clustering and Morphological Learning

John Mansfield,[a] Carmen Saldana,[b,c] Peter Hurst,[a] Rachel Nordlinger,[a] Sabine Stoll,[b,c] Balthasar Bickel,[b,c] Andrew Perfors[d]

[a]*School of Languages and Linguistics, University of Melbourne*
[b]*Department of Comparative Language Science, University of Zurich*
[c]*Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich*
[d]*School of Psychological Sciences, University of Melbourne*

Received 2 February 2021; received in revised form 18 October 2021; accepted 22 January 2022

Abstract

Inflectional affixes expressing the same grammatical category (e.g., subject agreement) tend to appear in the same morphological position in the word. We hypothesize that this cross-linguistic tendency toward *category clustering* is at least partly the result of a learning bias, which facilitates the transmission of morphology from one generation to the next if each inflectional category has a consistent morphological position. We test this in an online artificial language experiment, teaching adult English speakers a miniature language consisting of noun stems representing shapes and suffixes representing the color and number features of each shape. In one experimental condition, each suffix category has a fixed position, with color in the first position and number in the second position. In a second condition, each specific combination of suffixes has a fixed order, but some combinations have color in the first position, and some have number in the first position. In a third condition, suffixes are randomly ordered on each presentation. While the language in the first condition is consistent with the category clustering principle, those in the other conditions are not. Our results indicate that category clustering of inflectional affixes facilitates morphological learning, at least in adult English speakers.

Author contributions: JM: Conceptualisation, Methodology, Formal Analysis, Data Curation, Writing – Original Draft, Project Administration. CS: Conceptualisation, Methodology, Software, Validation, Formal Analysis, Visualisation, Writing – Original Draft. PH: Software. RN: Conceptualisation, Methodology, Writing – Review and Editing, Funding Acquisition. SS: Conceptualisation, Writing – Review and Editing. BB: Conceptualisation, Methodology, Formal Analysis, Writing – Review and Editing. AP: Conceptualisation, Methodology, Resources, Writing – Review and Editing, Funding Acquisition.

Correspondence should be sent to John Mansfield, School of Languages and Linguistics, University of Melbourne, Babel Building, Parkville, VIC 3010, Australia. E-mail: john.mansfield@unimelb.edu.au

Moreover, we found that languages that violate category clustering but still follow fixed affix ordering patterns are more learnable than languages with random ordering. Altogether, our results provide evidence for individual biases toward category clustering; we suggest that this bias may play a causal role in shaping the typological regularities in affix order we find in natural language.

Keywords: Morphology; Learning biases; Category learning; Artificial language learning

1. Introduction

The principles that govern the ordering of affixes in the world's languages have been much discussed in the literature from many theoretical perspectives (for a summary, see Manova, 2015). Despite great diversity in the morphological organization across languages, some types of affix configurations occur much more frequently than others. A common pattern, and one that is assumed as the default in many theoretical approaches, is that affixes expressing the same grammatical category tend to appear in the same morphological position in the word. This consistent positioning of affixes according to the grammatical category has been labeled "category clustering" (Mansfield, Stoll, & Bickel, 2020; see also Crysmann & Bonami, 2016; Good, 2016). In this study, we investigate the role that cognitive mechanisms involved in language learning might play in shaping this affix ordering preference cross-linguistically.

Previous research using artificial language learning (ALL) experiments has shown that language learners' preferences often mirror typologically frequent word and affix orders, thus suggesting a causal link between cognition and cross-linguistic regularities (Culbertson & Adger, 2014; Culbertson, Smolensky, & Legendre, 2012; Fedzechkina, Jaeger, & Newport, 2012; Hupp, Sloutsky, & Culicover, 2009; Maldonado, Saldana, & Culbertson, 2020; Saldana, Oseki, & Culbertson, 2021; Tabullo et al., 2012). Several studies show that learners tend to regularize word-order variation in the input, reflecting a preference for consistent ordering of words and word types within phrases (Culbertson & Newport, 2015; Culbertson et al., 2012; Fehér, Wonnacott, & Smith, 2016; Saldana, Kirby, Truswell, & Smith, 2019). We might expect that the consistent ordering of words by grammatical category will be reflected in morphology as consistent ordering of affix categories. In natural languages, morphological composition often mirrors the syntactic and semantic composition of elements (Bybee, 1985; Manova & Aronoff, 2010), which has led to theories that morphology is driven by the same fundamental mechanisms as syntax (e.g., Baker, 1985; Foley & Van Valin 1984). ALL experiments focusing on morphology have shown a preference for orderings that reflect syntactic and semantic composition, that is, where linear adjacency between affixes is determined by the order in which they compose with each other (Maldonado et al., 2020; Saldana et al., 2019, 2021). However, although these results are suggestive of a bias toward category clustering, there is hitherto no direct experimental evidence investigating whether this bias is present in morphological learning.

In the present study, we use ALL techniques to assess the link between individuals' biases in language learning and the predominance of category clustering cross-linguistically. We find that learners achieve more accurate suffix learning when they are taught words that conform to

morphological category clustering, compared to learners taught words that violate clustering. The magnitude of this bias appears to be relatively modest. Our results also allow for comparison between two grammars with different types of clustering violation, that is, between a grammar where morpheme order is rule-governed but unclustered and another with free morpheme order. Here, we find that the type of violation more frequently attested in natural languages (i.e., rule-governed unclustered) is more easily learned than the rarely attested type (i.e., free morpheme order).

2. Category clustering in inflectional morphology

Linear ordering rules are productive means of encoding differences in semantic composition in most or all languages. For example, the linear order of the constituents in the English sentence "*The cat chased the dog"* makes it unambiguous that *the cat* is the chaser and *the dog* the chasee; these roles are reversed when the constituent order is flipped in "*The dog chased the cat*." Linear order can also encode differences in semantic composition in morphology. For example, person/number agreement affixes on Swahili verbs index the subject participant in one position and object participant in another position (1).

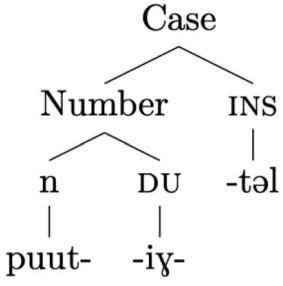
Swahili
a. ni-li-wa-lipa
1SG.S-PST-3PL.O-pay
"I paid them."
b. wa-li-ni-lipa
3PL.S-PST-1SG.O-pay
"They paid me." (Adapted from Stump, 1993) (1)

In other instances, linear ordering may be flexible without affecting semantic composition. Such flexibility is often said to occur more in syntax than in morphology (e.g., Anderson, 1992, p. 261; Jackendoff & Audring, 2019, p. 20). For example, affixes in Spanish (e.g., *bonit-o-s* "beautiful-MASC-PL") do not allow any permutation, while adjectives within a given noun phrase can appear before or after the noun (e.g., *bonitos poemas* "beautiful poems," *poemas bonitos* "beautiful poems").

Generalizing across all elements of a given grammatical category is fundamental to linguistic analysis, and these generalizations produce statements of linear order, such as subject–verb–object (SVO) constituent order or number-case affix order. While there are various syntactic, semantic, and pragmatic factors that may change the order in certain situations in a given language, these factors are expected to operate independently of specific V and O lexemes. Thus, syntactic theory represents constituents and their linear ordering in terms of grammatical categories rather than specific lexical items (Good, 2016, p. 55). Similarly, in morphological theory, the ordering of affixes is usually represented using grammatical categories, implying that by default, all affixes of the same category should appear in the same position. In theories such as Distributed Morphology (Halle & Marantz 1993), affixation is driven by syntax; therefore, affixes of the same category should appear in the same position

since this reflects a syntactic category node. For example, the word structure stem–number–case in the Mansi language is interpreted as a direct reflection of an underlying syntactic node structure (2), producing a consistent category ordering across individual affix values (compare 2a and 2b; Embick & Noyer 2001, p. 559).

Mansi



a. puut-iɣ-təl
pot-DU-INS
"by means of two pots";
b. puut-ət-nəl
pot-PL-ABL
"from many pots" (Keresztes, 1998, p. 410; Embick & Noyer, 2001, p. 559). (2)

Other theories such as Paradigm Function Morphology (Stump, 2001) treat affixation as fundamentally autonomous from syntax, but in this theory, the paradigmatic alternations between different feature values (e.g., 1SG vs. 2SG subject agreement) are the driving force behind affix positioning. This approach captures the category-based ordering of affixes that realize these paradigmatic alternations, for example, the consistent S-tense-O-stem ordering in Swahili (3).[1]

Swahili
a. tu-li-ku-lipa
1PL.S-PST-2SG.O-pay
"We paid you."
b. wa-me-tu-lipa
3PL.S-PERF-1PL.O-pay
"They have paid us."
c. ni-ta-wa-penda
1SG.S-FUT-3PL.O-like
"I will like them" (Stump, 1993; citing Gleason, 1955). (3)

The consistent positioning of affixes according to the grammatical category has been labeled "category clustering" (Mansfield et al., 2020; see also Crysmann & Bonami, 2016; Good, 2016). However, although category clustering is fundamental in theories of syntax and morphology, it is not difficult to find violations of the principle. A syntactic violation can be found in French noun–adjective ordering, which like the Spanish example above exhibits

some flexibility but also has fixed orderings for specific adjectives, at least partly determined by semantics (Bouchard, 1998; Waugh, 1977). Morphology arguably presents more frequent clustering violations. For example, in some languages, affixes of the same category (i.e., same morphological feature) appear in different positions depending on their specific combinations of feature values. This can be seen in the ordering of S and O verb agreement suffixes in the Niger-Congo language Fula (4): O agreement is peripheral to S agreement in most forms (4a), but S agreement is realized peripherally to O agreement in those forms in which 1SG.S coincides with 2SG.O or 3SG.O (4b).

Fula
a. mball-u-ɗaa-mo'
help-REL.PST-2SG.S-3SG.O
"You helped him."
b. mball-u-moo-mi'
help-REL.PST-3SG.O-1SG.S
"I helped him" (Arnott, 1970; Stump, 2001, p. 151). (4)

Free ordering produces another type of clustering violation, and although as mentioned above, it is said to be rare in morphology, there are several attested examples (Bickel et al., 2007). The Australian Aboriginal language Murrinhpatha provides one such example, with flexible ordering of some tense and number suffixes without any change in meaning (5).

Murrinhpatha
a. purne-lili-dha-nime
go.3PAUC.PST-walk-PST-PAUC.M
b. purne-lili-nime-dha
go.3PAUC.PST-walk-PAUC.M-PST
both: "They (paucal, masculine) were walking" (Mansfield, 2019, p. 160). (5)

A recent study investigating category clustering in verbal paradigms shows that despite the violability of the principle, there is nevertheless evidence for a universal, probabilistic clustering bias in verbal agreement markers (Mansfield et al., 2020). This study investigated agreement paradigms in 136 languages and found that some degree of clustering violation is relatively common: half of the S and O verbal agreement paradigms had members of the paradigm spread across multiple morphological positions. However, even where clustering was not absolute (i.e., not all members of a paradigm in the same position), markers belonging to the same paradigm still showed a strong tendency to occur in the same position. The authors further undertook a corpus study of free prefix ordering in the Kiranti language Chintang. This showed that although any order of prefixes is grammatical, speakers exhibited a probabilistic bias toward category clustering, tending to place S agreement markers in one position and O agreement markers in another position. The authors hypothesize that this reflects a cognitive bias toward category clustering in morphology, though they note that the bias remains to be investigated in categories other than S/O agreement (Mansfield et al., 2020, p. 273).[2]

The prevalence of category clustering in natural languages and the assumptions made in linguistic theory suggests that there may be a principled bias favoring structures that conform

to category clustering. To date, no experimental studies have tested this conjecture. In the current study, we focus on a possible learning bias that may favor clustering in inflectional morphology. Specifically, we use ALL techniques to test whether participants learning a novel miniature language find it easier to learn its morphology when it complies with category clustering principles—that is, when all affixes realizing the same category appear in the same position.

2.1. Evidence from ALL studies

ALL is a widely used technique for investigating the preferences and biases learners bring to a language-like task, with the potential to provide insight into natural language learning. ALL experiments invite participants to learn and reproduce small sets of words or phrases, on the assumption that this taps into the same cognitive and learning principles at work in language, even though the small vocabulary, brief timescale and experimental setting are all very different from the learning processes in actual language use.

Studies investigating the learnability of probabilistic word order variation show that participants regularize input variability during learning and use (Culbertson & Newport, 2015; Culbertson et al., 2012; Fehér et al., 2016; Wonnacott, Newport, & Tanenhaus, 2008). In these studies, participants are taught synonymous word order patterns whose occurrence is probabilistic and not conditioned on any aspect of their linguistic or extralinguistic context. When tested on their own production, participants tend to regularize inconsistencies in the input, that is, to make word ordering more consistent.

There is also evidence that regularization can be modulated by the nature of the specific structures being learned. For example, learners regularize harmonic word order patterns (i.e., which are either consistently head-initial or head-final) more than non-harmonic patterns within the noun phrase (Culbertson & Newport, 2017; Culbertson et al., 2012; Culbertson, Franck, Braquet, Barrera Navarro, & Arnon, 2020). At the sentence level, Wonnacott et al. (2008) show that learners have a strong preference for word orders in which the agent precedes the patient (i.e., for VSO orders over VOS orders). In both of these cases, the biases uncovered match typological generalizations in word order patterns: Harmonic orders within the noun phrase are most frequent among the languages of the world (Greenberg, 1963), and so are constituent order patterns where the subject precedes the object (Dryer, 2013a). This suggests that despite the artificiality of the experimental tasks, they may reveal cognitive biases that are also operative in natural language learning and transmission.

Learning biases toward specific ordering patterns have also been investigated in the domain of morphology using ALL. Hupp et al. (2009) found that, in accordance with the cross-linguistic suffixing preference (Bybee, Perkins, & Pagliuca, 1994; Dryer, 2013b; Himmelmann, 2014), speakers of a suffixal language such as English are more likely to treat two words as referring to the same referent if they differ in their endings rather than in their beginnings. However, Martin and Culbertson (2020) show that this preference is not present in speakers of prefixal languages such as Kîîtharaka (Atlantic-Congo, Kenya), which suggests that learning biases are mediated by prior linguistic knowledge.

Recent ALL experiments have further revealed a learning bias toward specific affix orderings within multi-affix words. Saldana et al. (2021) taught native English speakers an artificial language with noun stems and accusative case and plural number affixes. The input language indicated only whether each affix preceded or followed the noun stem but did not provide any information about the relative order of case and number affixes (instances of plural accusatives were held back during training). Learners consistently produced numbers closer to the noun stem than the case when asked to produce these previously unseen plural accusatives. As before, learners' preference matched a cross-linguistic tendency in which number markers tend to be ordered closer to noun stems than case markers (Bybee, 1985; Foley & Van Valin, 1984; Greenberg, 1963). This result adds to the growing body of work showing the existence of a (second-language) learning bias toward compositionally transparent orders, that is, toward the linearization of meaningful elements that match their order of composition (for evidence in the word-order domain, see Culbertson & Adger, 2014; Martin, Ratitamkul, Abels, Adger, & Culbertson, 2019; for evidence in person-number morphology, see Maldonado et al., 2020).

Further evidence linking learning biases to more general linguistic features comes from experiments where artificial languages are transmitted across several generations of learners. Kirby, Cornish, and Smith (2008) and Kirby, Tamariz, Cornish, and Smith (2015) developed an iterated ALL paradigm to explore the impact of learning biases over multiple iterations, with learners becoming "teachers" to new generations of learners. The studies demonstrate the emergence of compositional linguistic structure from an initial unstructured language as it is transmitted down generations of participants organized in transmission chains in the lab. The initial input for these experiments is single-word labels with no internal structure but modifications introduced through imperfect learning and transmission lead to the emergence of "morphology," that is, word parts associated with simple meanings (in this case, color, motion, or shape features of an image). These simple meanings are then further combined to create complex meanings (e.g., "a red circle moving in a straight line"). Crucially, the emergence of such a structure is dependent on the presence of both a pressure for learnability (imposed during transmission) and a pressure for expressivity (imposed during communicative interaction, see Kirby et al., 2015). A more recent experiment required participants to learn the output languages from iterated learning experiments and showed that the output languages with more systematic morphology are learned more accurately than those with less morphology (Raviv, Kloots, & Meyer, 2021).

The iterative emergence of morphology was replicated with a more adequate sample size in Beckner, Pierrehumbert, and Hay (2017) and with a more complex compositional structure in Saldana et al. (2019). In these experiments, morphological structure emerges relatively quickly, and importantly to our study, category clustering occurs as well. For example, one of the final languages that evolved in Kirby et al. (2008), exp2, Chain 4) predominantly expresses color (black, blue, or red) in a prefix position, and motion (straight line, zig-zag, or circular) in a suffix position. Again, in the follow-up study by Beckner et al. (2017), much of the emergent morphology conforms to category clustering as in (6) where similar phonological

strings emerge in "stem position" and "suffix position," though these strings do not always map neatly to semantic properties.

a. sancla-v "one green berry"
b. sanklo-ki "two green berries"
c. siko-ven "one blue phone"
d. suki-ki "two blue phones"
(Beckner et al., 2017; diffusion chain 4; hyphens added for clarity). (6)

Although these studies rarely show an emergence of unclustered morphology, it does occur at times. For example, the set of words in (7), which are taken from one of the languages in Beckner et al. (2017), suggests that a stem form *tso* "berry" has emerged, with number and color features expressed via suffixes (7a, 7b) or prefixes (7c, 7d). This suggests that participants who spontaneously develop morphological structure are capable of learning and using unclustered morphology.

a. tso-vol "two blue berries"
b. tso-kiki "three blue berries"
c. de-tso "two red berries"
d. tri-tso "three red berries"
(Beckner et al., 2017; diffusion chain 5; hyphens added for clarity). (7)

Taken together, the aforementioned studies suggest that artificial language learners prefer word and affix orders where elements of the same category appear consistently in the same position; however, there is hitherto no study providing direct evidence for such a preference. The previous studies focus on the importance of compositional morphology itself but do not test the role of linear ordering in learning compositional structure. In the present study, we address this gap, by testing whether the hypothesized learning bias for category clustering in morphology can be shown to exist in an ALL task. In doing so, we do not claim to replicate the natural language learning and transmission behaviors that by hypothesis led to the prevalence of clustering in S/O agreement. Nonetheless, if we find evidence for a clustering bias in this language-like experimental task, it is reasonable to suppose that this same bias has a role in shaping natural languages when they evolve over generations of learners. A broader question is whether the clustering bias is specific to language or whether it is a domain-general principle that might also show up in a wider range of learning tasks; however, this complex question is beyond the scope of the current study.

2.2. *Hypothesis: Category clustering in artificial language learning*

Following the lead of previous studies that show cross-linguistic regularities to be mirrored in experimental learning biases, we use an ALL paradigm to test a bias toward category clustering. We formulate our hypothesis as follows:

> When participants are taught an artificial language with inflectional suffixing structure, they will more accurately identify the correctly suffixed forms in a grammar that

conforms to category clustering, compared to non-clustered grammars of either of the following types:

(a) A grammar with fixed positions for each suffix, but different positions for different suffixes of the same category (as in Fula (4) above);
(b) A grammar with a free variation of suffix ordering (as in Murrinhpatha (5) above).

We were agnostic as to whether there would be a further disparity between the two types of clustering violation, though we address this issue in our discussion (Section 5). The experiment described below was in fact our second ALL experiment on category clustering. We previously ran a slightly different experiment in which a more randomized design returned more complex results. Details of the earlier experiment are included as Supplementary Materials E.

3. Method

We taught adult English speakers a miniature artificial language containing morphologically complex labels for different objects.[3] The objects varied in shape, color, and number; their corresponding labels were composed of noun stems as well as two suffixes referring to color and number features. At the test, participants were asked to recall the labels they were taught but also to generalize to new items. Our experimental setup drew inspiration from ALL studies mentioned in Section 2.1 but was not directly modeled on any previous experiment.

3.1. Participants

We recruited 364 English native speakers through the Amazon Mechanical Turk platform.[4,5] We targeted adult English speakers because this allowed us to access a large pool of potential participants, even though this somewhat limits the generality of our results, as the language experience of participants may influence their learning behavior with the artificial language (Martin & Culbertson, 2020). We consider the implications of using English speakers in the discussion below (Section 5.3). Participants were paid US$5 for an experimental session that took up to 25 min to complete. We applied experimental features to periodically refocus attention (see Hauser & Schwarz, 2016; Mellis & Bickel, 2020) and applied a pre-registered exclusion on participants whose response times suggested inattentive clicking (see Supplementary Materials A), resulting in 297 participants being included in the analysis.

3.2. Input language

Both training and testing regimes used text on screen (as opposed to auditory stimuli). The lexicon in the input language contains noun labels for four different shapes (cross, circle, wave, star), which always appear in one of three colors (red, green, blue) and one of three numbers (singular, plural, dual). Thus, there are $4 \times 3 \times 3 = 36$ visual stimuli represented in the language. This meaning space was similar to those used in previous ALL experiments

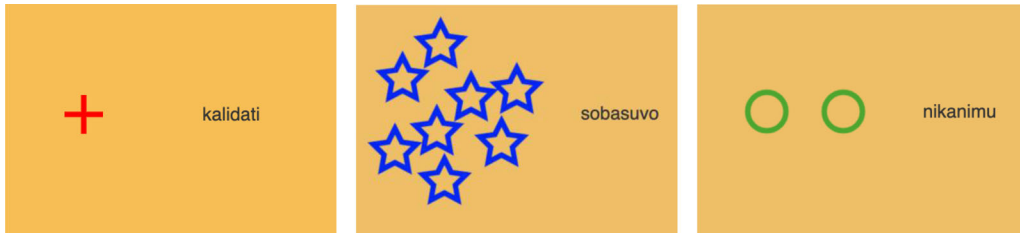


Fig 1. Examples of shapes and names presented in the experiment.

(e.g., Beckner et al., 2017; Kirby et al., 2008; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010). We selected color and number as semantic features to be encoded by our suffixes because they can be clearly represented in simple images. The number feature is similar to natural languages, in which number-marking inflections are widely observed. The color feature is not typically an inflectional feature in natural languages,[6] and in this respect, our experiment is only approximately language-like.

Nouns contain a disyllabic lexical stem that corresponds to one of the four shapes, as well as two monosyllabic suffixes, one corresponding to color and the other to number. We used exclusively suffixing morphology since adult English speakers have previously been shown to learn suffixes more readily than prefixes (St Clair, Monaghan, & Ramscar, 2009; Hupp et al., 2009). All syllables are CV, and the disyllabic stems never contain a syllable that matches a suffix syllable. The lexicon of stems and suffixes was as follows:

Shape stems: *kali* "cross," *nika* "circle," *fugi* "wave," *soba* "star"[7]
Color suffixes: *-da* "RED," -mu "GREEN," *-vo* "BLUE"
Number suffixes: *-ti* "SG," *-ni* "DUAL," *-su* "PL"

Fig. 1 illustrates three example stimuli containing different form-meaning mappings. Glosses in (8) show the morphological structure of these examples, which was *not* taught explicitly.

a. kali-da-ti
cross-RED-SG
b. soba-su-vo
star-pl-BLUE
c. nika-ni-mu
circle-DUAL-GREEN. (8)

3.3. Conditions

The three conditions varied according to three different "grammars" determining the positioning of affixes. Each participant was randomly assigned to one of the three grammars, and after the exclusions mentioned above, the number of participants was 97 simple, 106 complex, and 94 random. The grammars are defined as follows, where Slot 1 is the position immediately to the right of the stem, and Slot 2 is the position to the right of Slot 1:

Simple grammar: Morphological features are assigned to slots, with color in Slot 1 and number in Slot 2. This grammar has a consistent morphological template of the form stem–color–number. This grammar corresponds to fully clustered, canonical systems of inflection (Baerman & Corbett, 2012; Stump, 2015, Chap. 2).

Complex grammar: Slot assignment is conditioned on specific suffix combinations. Number is placed in Slot 1 when its value is plural or when the adjacent color value is blue (i.e., -PL-RED, -PL-GREEN, -PL-BLUE, -SG-BLUE, -DUAL-BLUE); otherwise, number is placed in Slot 2 (i.e., -RED-SG, -RED-DUAL, -GREEN-SG, -GREEN-DUAL). This means that RED, GREEN, SG, and DUAL may each appear in either slot, while PL and BLUE appear consistently in Slots 1 and 2, respectively. This gives four pairings in which color precedes number, and five in which number precedes color. This grammar corresponds to the type of clustering violation exemplified by Fula (4 above).

Random grammar: The grammar does not determine suffix ordering, and the order is randomly assigned on each presentation of a label, with each ordering being of equal probability. For example, a single red wave can be labeled as either wave-RED-SG or wave-SG-RED. This grammar corresponds to the type of clustering violation exemplified by Murrinhpatha (5 above).

Our research hypothesis is that participants should find it easier to learn the meaning of affixes in the simple grammar (with consistent category clustering in affix order), compared to the non-clustered grammars, either complex or random.

3.4. Procedure

The experiment was conducted online via the participants' individual computer terminals; all instructions were provided in English. Participants were told that they would be learning a new language, but they were not given any further information about the structure of the language or the experimental hypothesis.[8] The session was divided into three phases of interleaved training and testing blocks.

3.4.1. Training blocks

Participants were first exposed to a series of images paired with their corresponding labels as in Fig. 1. At each trial, they were asked to learn a single image paired with its label, which was displayed until the participant pressed a button to move onto the next presentation. Through three blocks of training, participants were taught 24 image labels, a subset of the total 36 labels in the language. In Training block 1, they were exposed to eight labels. In Training block 2, they were presented eight novel labels as well as the same eight labels they were already taught in the first block. Finally, in training block 3, they were exposed to a further eight new labels as well as the previous 16. Training block 1 showed each label in the block twice, while Blocks 2 and 3 showed each image in the block just once so that the

training blocks were of relatively similar length. Thus, the number of presentations in each block was as follows:

Training block 1: $8 \times 2 = 16$
Training block 2: 16 (incl. re-teaches) $\times$ 1 = 16
Training block 3: 24 (incl. re-teaches) $\times$ 1 = 24

The training regime thus provides a total of four presentations for those stimuli allocated to Training block 1, two presentations for those in Block 2, and one presentation for those in Block 3.

The allocation of images to training blocks was structured so that blocks included paradigmatic alternations between words differing by a single suffix, for example, *kali-da-ti* "cross-RED-SG" and *kali-vo-ti* "cross-BLUE-SG." The eight labels taught in Training block 1 included two different inflectional forms for each of the four stems. The suffix alternations used were RED vs. BLUE, and SG vs. PL, with the GREEN and DUAL suffixes held back until Training block 2. Training block 2 expanded the set of labels to include four inflectional forms for each stem, and Training block 3 included six forms for each stem (for full details of block composition, see Supplementary Materials B).

3.4.2. Testing blocks

Participants were tested on their ability to recall the form-meaning mapping they had been trained on as well as their ability to generalize the learned forms to novel meanings. On each test trial, participants saw an image of a shape along with an array of four possible labels to describe it (the target form and three distractors, presented in random order on the screen) and were asked to select the correct label.

Testing was administered in three blocks, one after each training block. The first and second test blocks contained one test trial for each label that had been taught up to that point. The third test block comprised one test trial for each label taught, as well as one test trial for each of the 12 labels (out of the possible 36) that were held back during training. Thus, the third testing block included tests that specifically required learners to generalize or extrapolate the morphology to unseen images. In order to succeed in trials with unseen meaning, participants had to learn stems and affixes individually, that is, learn the compositional structure of the labels rather than rote-learn them. The number and type of tests in each block are as follows:

Test block 1: eight seen
Test block 2: 16 seen
Test block 3: 24 seen + 12 unseen
TOTAL: 48 seen + 12 unseen

Test distractors were selected in such a way that suffix learning was required to identify the correct answer. The three distractors all had the correct stem, but one or both suffixes were incorrect. For example, in (9), knowledge of both suffixes is required in order to identify the correct answer.

Suffix test
a. kali-da-ti
cross-RED-SG
b. kali-vo-ti
cross-BLUE-SG
c. kali-da-su
cross-RED-PL
d. kali-vo-su
cross-BLUE-PL. (9)

These distractors were designed to test suffix *learning* but did not require the participant to have learned a system of suffix *ordering*. Only one form in each testing array had the correct suffixes, so it was theoretically possible to attain 100% accuracy based only on acquiring suffix meanings, without learning any ordering rules at all. In this way, we used the three morphological grammars as conditions for learning the meaning of suffixes, but the ordering systems were not in themselves the learning targets.

The distractor format above was applied to the testing regime as a whole, yielding 60 "suffix tests" for each participant. In addition, to check whether participants were also learning stems, we added eight auxiliary "stem tests" (four in Training block 1, two in each of the other blocks), where distractors all had correct suffixes but incorrect stems.

Stem test
a. kali-da-ti
cross-RED-SG
b. nika-da-ti
circle-RED-SG
c. fugi-da-ti
wave-RED-SG
d. soba-da-ti
star-RED-SG. (10)

4. Results

The suffix-test data consist of 60 test responses for each of 297 participants, giving 17,820 test responses in total. Fig. 2 shows participants' accuracy scores (proportion of answers that are correct) aggregated across test phases and grouped by grammar. In line with our hypothesis, a visual inspection of our results suggests that participants with the simple grammar score the highest, followed by those with the complex and random grammars. The results also suggest that participants with the complex grammar perform slightly better than those with the random grammar.

Fig. 3 illustrates the development of these differences across test blocks. In all three grammars, mean test accuracy decreases as participants progress through the blocks, which may reflect the incremental demand on memory built into our design. Responses in the complex

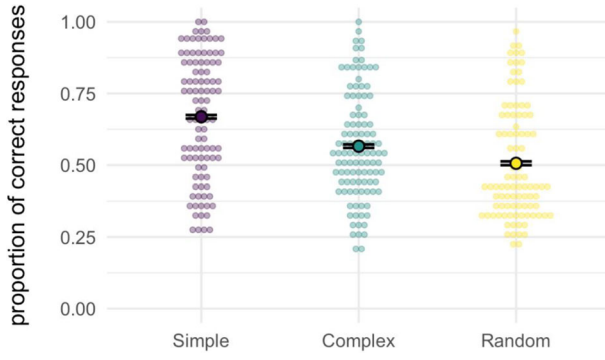


Fig 2. Participants' scores on suffix test trials aggregated across the three test phases and grouped by grammar. Shaded dots represent individual scores, and circled dots depict the groups' means and standard errors.

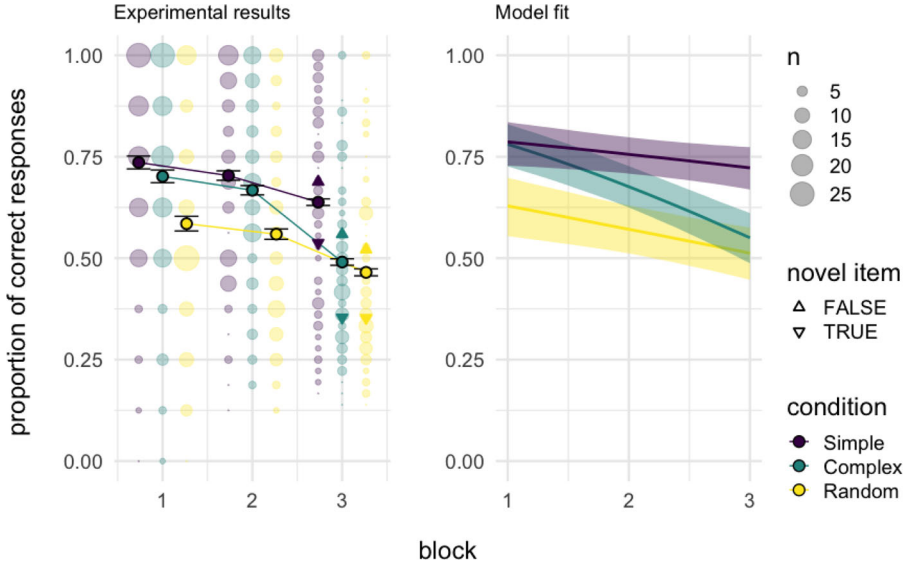


Fig 3. Mean participant accuracy score by grammar and test block. Left: Empirical data obtained from the experiments. Shaded circles represent individual scores, and larger circles represent more individuals. Filled circles represent mean accuracy scores, and the error bars represent the standard error. At Block 3, the mean is calculated across both novel and familiar items together; the mean scores split by novel versus familiar items are additionally illustrated by the upward (seen items) and downward (unseen "novel items") triangles. Right: Model estimates of the fitted Bayesian beta-binomial model for the fixed effect of the block. Thick lines represent the predicted accuracy means conditioned on grammar and block. The shaded area shows the 95% uncertainty intervals.

grammar are similar to the simple grammar in Test blocks 1 and 2, but when participants reach Test block 3, the complex grammar response accuracy degrades considerably. Responses in the random grammar are less accurate across all test blocks, with the exception of the third block, where accuracy scores in the random grammar are comparable to those in the complex grammar.

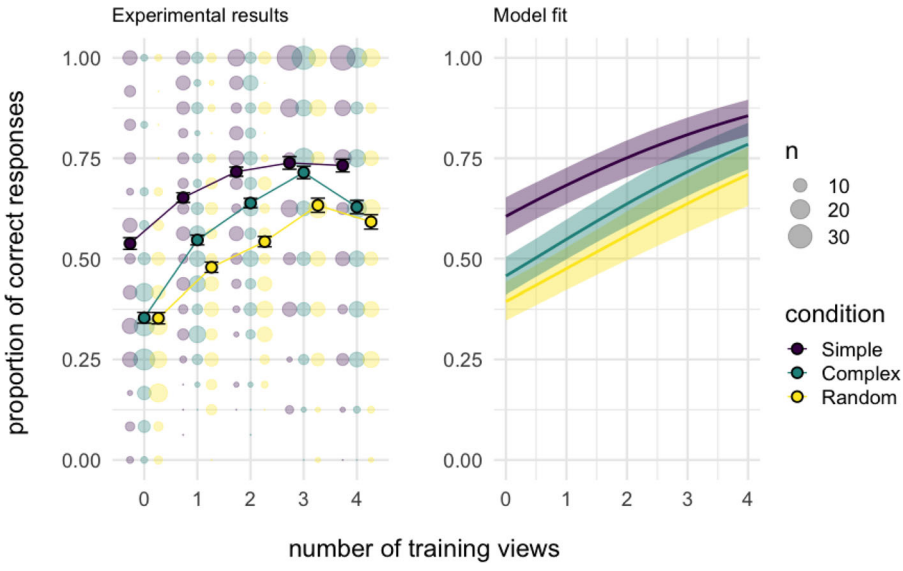


Fig 4. Number of training views. Left: Empirical data obtained from the experiments. Shaded circles represent individual scores, and larger circles represent more individuals. Filled circles represent mean accuracy scores, and the error bars represent the standard error. Right: Model estimates of the fitted Bayesian Beta binomial model for the training views fixed effect. Thick lines represent the predicted accuracy means conditioned on experimental condition and block. The shaded area shows the 95% uncertainty intervals.

As outlined above, the training regime is incremental, and items taught earlier are repeated in later blocks. Even though participants' overall test scores might decrease as they attempt to learn a larger set of items, we nevertheless expect accuracy on individual items to increase as a function of previous exposure: The more exposure to an item, the better the recall. Fig. 4 illustrates mean accuracy scores grouped by grammar and the number of times that the target item has been viewed in training. The unseen items in the third testing block are represented here as having zero training views.[9] The unseen items produce the lowest average scores, with the simple grammar average being higher than the complex and random grammars. The figure shows a monotonic improvement in the simple grammar as training views increase, but for complex and random grammars, this improvement is maintained only up to three views but not a fourth. Note that testing on an item viewed four times occurs only in the third testing block, which as we saw above exhibits generally lower accuracy for the complex and random grammars. Figs. 3 and 4 together suggest that early-viewed items were learned better after repeated viewing, but as the number of different testing items increased, overall performance decreased.

We used a Bayesian mixed-effects regression model to estimate the probability of our category clustering hypothesis (simple > complex & random). We also estimated the probability that one type of non-clustered grammar is more learnable than the other (complex > random) as suggested by our data visualization. The mixed-effects model allows us to analyze the main effects and their interactions, while also taking into account the individual

variability of participants. Using the *brms* package (Bürkner, 2017, 2018) as an R (R Core Team, 2013) interface to Stan (Stan Development Team, 2020), we ran a beta-binomial Bayesian regression model[10] predicting participants' performance by grammar (simple, complex, and random), testing block (three blocks, with the intercept at Block 1), and the number of training views (which vary between 0 and 4, with the intercept at 0 views). We also tested for interactions between grammar and block and between grammar and the number of views. The categorical predictor grammar was coded to compare complex to random and compare simple to the average of the two (i.e., "reverse Helmert" coding, Wendorf, 2004); the intercept is the grand mean of all three grammar conditions. As random effects, we included intercepts for subject as well as by-subject slopes for the effects of views and block. The model formula is shown in (11) (in R-style format). We set the same Student-t prior on all fixed effects as well as on the intercept (DF $= 6$, $\mu = 0$, $\sigma = 1.5$); for the random effects, we set a half-Cauchy prior with scale parameter 10; for the precision parameter ϕ, we set an exponential prior with a rate of 1 (McElreath, 2016, p. 348). The diagnosis of the model fit is included in Supplementary Materials C.

$$\text{correct responses} \mid \text{trials}(n) \sim \text{Grammar} + \text{Views} + \text{Block} + (\text{Grammar:Views}) + (\text{Grammar:Block}) + (1+\text{Views}+\text{Block}|\text{Subject}). \quad (11)$$

The structure of the model reported here differs from the preregistered model: The latter did not include the number of training views as a factor but instead contained novelty (seen vs. unseen items during training), which only applied to the third block of training as described above. Results are very similar across models. However, including the number of training views as a predictor further allows us to observe whether accuracy scores increase as a function of training. We report the pre-registered model in the Supplementary Materials D. The scripts and results for both the pre-registered model and the model reported here are available at the data repository for this study.[11]

Fig. 5 shows the model's posterior distribution densities for all fixed effects along with their means (solid black line) and 95% uncertainty intervals[12] (dashed gray lines). The results suggest that participants with the complex grammar have higher accuracy than those with the random grammar ($\beta = 0.34\,[0.15,\ 0.54]$, $SE = 0.10$), and in line with our hypothesis, participants with the simple grammar have slightly higher accuracy than the average of the complex and random grammars together ($\beta = 0.14\,[0.02,\ 0.26]$, $SE = 0.06$). The probability of our hypothesis (simple > complex & random), given our model, priors, and data is 0.99. The probability that (complex > random) is 1. The model further suggests that accuracy increases with the number of training views ($\beta = 0.35\,[0.30,\ 0.39]$, $SE = 0.02$) but decreases by block as the number of labels to be memorized increases ($\beta = -0.33\,[-0.41,\ -0.22]$, $SE = 0.05$). While the increase in accuracy with training views is similar across grammars (Complex vs. random, $\beta = 0.02[-0.03,\ 0.07]$; simple vs. complex & random, $\beta = 0.00\,[-0.03,\ 0.03]$), the decrease in accuracy between blocks is not: the proportion of correct responses decreases more in complex than in random ($\beta = -0.15\,[-0.25,\ -0.04]$, $SE = 0.05$), while in Simple it decreases slightly less than in complex and random together ($\beta = 0.07[0.01,\ 0.13]$, $SE = 0.03$).

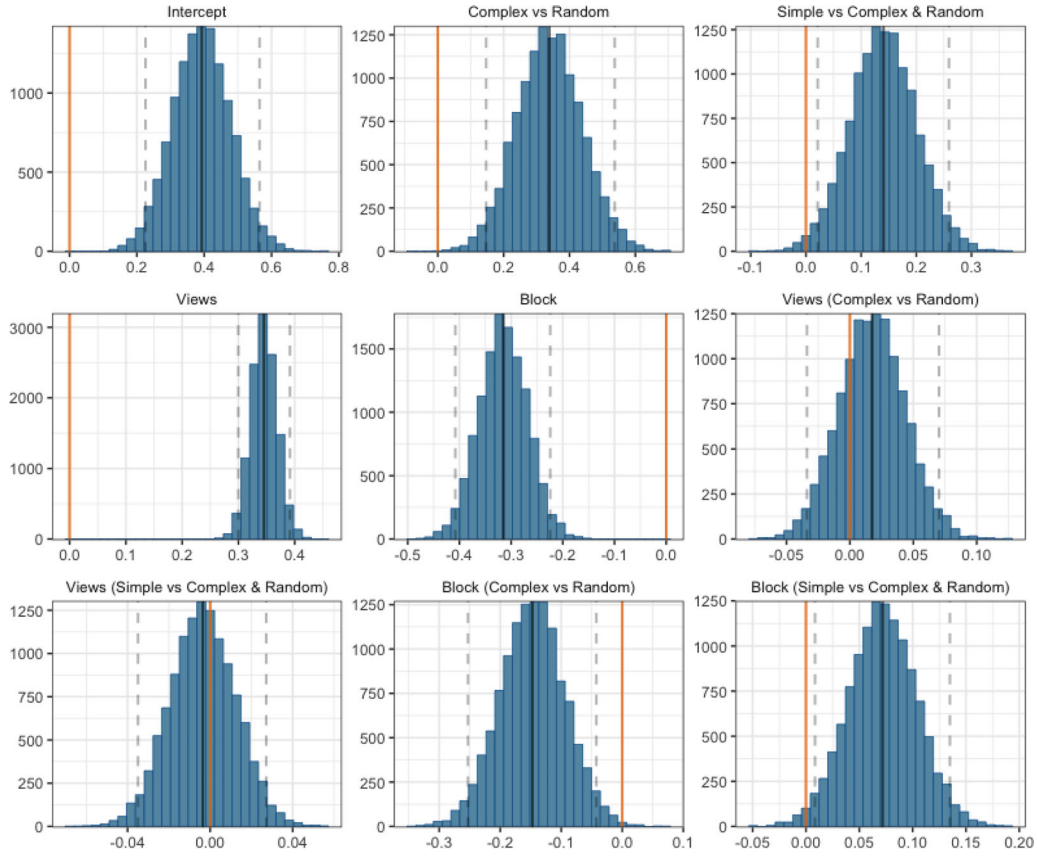


Fig 5. Model's posterior distribution densities for all fixed effects and their interactions, along with their point mean estimates (solid black line) and 95% uncertainty intervals (dashed gray lines).

As noted above, we ran a smaller number of auxiliary test trials focused on stem learning. These were analyzed separately as they are not critical to our hypotheses and thus not included in the main analysis. Fig. 6 illustrates participants' proportion of correct responses on stem test trials, grouped by grammar: accuracy scores were very high on stem tests across conditions.

5. Discussion

Our experimental results suggest that the random grammar produces less accurate affix learning than the complex grammar and that together these produce less accurate learning than the simple grammar. This supports our hypothesis that morphological category clustering, as exemplified in the simple grammar, is more favorable to affix learning than the two types of non-clustered grammar. However, the effect size of this main result ($\beta = 0.14$) is smaller

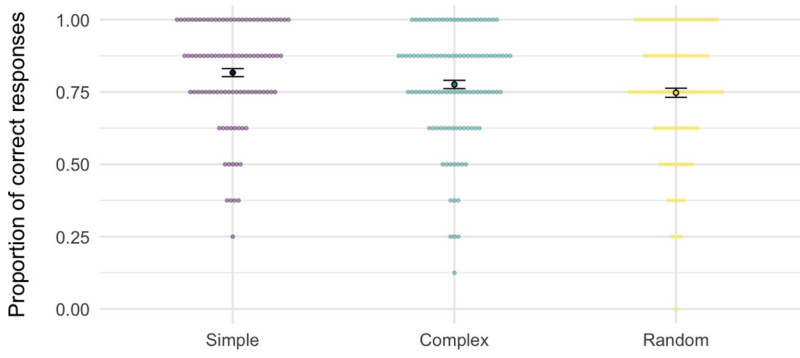


Fig 6. Participants' scores on stem test trials aggregated across the three test phases and grouped by grammar type. Shaded dots represent individual scores, and circled dots depict the groups' means and standard errors.

than the other fixed effects (training views $\beta = -0.33$; block progress $\beta = 0.35$), suggesting that the learning bias toward category clustering may be relatively moderate. We did not have a hypothesis about the relative learnability of complex and random types of non-clustering, but our results suggest that complex grammar is more favorable to affix learning than random grammar, and this difference is greater ($\beta = 0.34$) than the difference between clustered and unclustered grammars overall. The positive effect of complex, compared to random is driven by the first two test blocks, where complex achieved similar levels of accuracy to the simple grammar.

Our experiment also suggests an interaction between grammar and block progress, where block progress incrementally increases the number of labels to be memorized, and in the third block, participants are also tested on novel items. While block progress led to a decrease in accuracy across all grammars, this effect was smaller in the clustered (simple) grammar, compared to the other two; and between the two unclustered grammars, complex started off with higher accuracy scores and exhibited a steeper decrease during block progression. Accuracy scores for complex are similar to those for the simple grammar in Blocks 1 and 2, but they drop to a level similar to random by the third block. We also found that repeated views of the same lexical item improved its learning, and this effect was independent of the grammar; this result is expected given that learning is strengthened with repeated exposure to evidence.

5.1. Generalization and compositionality

The interaction between block progression and grammar suggests that there may have been differences in the learning strategies used by participants under different grammars. However, there are some ambiguities in the interpretation of these effects, as block progression involved both (a) an increased number of labels to be memorized and (b) testing on some novel items, requiring generalization, in the third block but not in preceding blocks.

The negative effect of block progression occurred to some extent under all grammars, but we also found grammatical differences, with greater degeneration occurring in the complex grammar, compared to random, and greater in these two together compared to simple. As

we argue below, both the increase in the number of labels tested and the requirement for generalization in Block 3 are problems that would seem more tractable using compositional learning, as opposed to whole-word learning. Therefore, the greater degeneration exhibited by participants in the complex grammar may reflect less compositional learning in this condition.

Our results suggest that participants deployed a combination of compositional and whole-word learning strategies in this experiment (cf. Marslen-Wilson, 2007). The positive effect of repeated training views suggests some degree of whole-word learning: otherwise, items trained just once in the second block (e.g., cross-BLUE-SG, cross-RED-PL) should have been no more difficult than thrice-trained items built from the same morphological elements (e.g., cross-RED-SG, cross-BLUE-PL). The fact that participants did better on individual items after repeated training suggests that whole-word learning played a part in their successful test responses. On the other hand, successful generalization to novel items in the third test block indicates some degree of compositional learning. Accuracy on these novel items appears to be higher in the simple grammar than the complex and random grammars, again suggesting that more compositional learning occurred in the simple grammar.

Since the novel items in Block 3 clearly required compositional learning, the greater reduction of accuracy under the complex grammar—especially the abrupt decrease in accuracy in Block 3—may partly reflect less compositional learning by these participants. In the complex grammar, each specific affix combination has a consistent form, but the paradigmatic categories of number and color are not consistently positioned, which may have made it more difficult for participants to segment these combinations into pairs of affixes. In the simple grammar, the categories were consistently positioned, which may aid segmentation. In the random grammar, categories were not consistently positioned, but since none of the affix combinations had a fixed form, this may have meant that whole-word learning was no more rewarding than compositional learning, resulting in lower accuracy in the random grammar for both trained and novel items.

The increased number of labels associated with block progression ($8 < 16 < 24$) may also have favored compositional learning. Compositional learning could be expected to show greater tolerance of proliferating labels, as long as these labels are composed of the same stems and affixes. In a fully compositional strategy, the learner only needs to learn a total of 10 items (four stems, three color suffixes, three number suffixes). But in a fully whole-word strategy, 24 items must be learned, that is, every trained item must be learned separately. Therefore, even if we interpret the negative effect of block progression as relating primarily to the memory load of proliferating labels, this is still compatible with the idea that participants in the complex grammar used less compositional learning.

Nonetheless, our finding of an interaction between block progress and grammar must be interpreted with caution. As mentioned above, block progress and the number of training views were not fully independent in our experimental design. Furthermore, the number of test trials with novel items was fairly small (eight per participant). A different experimental design would be required to fully investigate the balance of whole-word versus compositional learning.

5.2. Correspondence with typological distribution

The relatively weak bias for clustered over unclustered grammars appears to be consistent with the typological findings of Mansfield et al. (2020), which coded category clustering of verbal agreement markers in a sample of 136 languages from the typological database AUTOTYP (Bickel et al., 2017). This study did not implement a binary classification of languages as simple or complex but rather measured degrees of clustering based on individual affixes in each language, that is, the degree to which markers of the same category appear in the same position. Although the results showed an overall bias toward category clustering, the sample also revealed a substantial proportion of verbal agreement paradigms in which one or more markers are positioned inconsistently with the rest of the paradigm, even if the paradigm as a whole still shows more clustering than could be explained by chance. These paradigms thus have a structure that is to some degree complex, while remaining consistent with an overall bias toward category clustering. If there is indeed only a weak learning bias for simple over complex grammars, this might result in a cross-linguistic bias toward clustering but with many grammars showing a slight degree of complex affix positioning.

On the other hand, Mansfield et al. (2020) excluded random grammars from their sample since only two such grammars (Chintang and Bantawa) are attested in AUTOTYP. The distribution in AUTOTYP and other observations in the literature (Manova & Aronoff, 2010, p. 125; Jackendoff & Audring, 2019, p. 20) suggest random grammars are rare in the languages of the world. Since we found worse learning of the random grammar than the complex grammar, this is consistent with the hypothesis that learning biases disfavor random morphological positioning in comparison to complex morphological positioning.

The finding of complex > random grammar is also compatible with the literature exploring the acquisition of probabilistic linguistic variation: These studies show that conditioned variation is preferable to unconditioned variation (e.g., Samara et al., 2017; Wonnacott et al., 2017). For example, Wonnacott et al. (2017) taught adults and children a semi-artificial language where six English nouns were combined with two meaningless "particles." There were several experimental conditions, but most relevant here is one condition with lexically consistent noun-particle pairings for all six nouns, and another condition where just two nouns were consistently paired with particles, and the other four nouns were paired inconsistently with the two particles in a 50:50 ratio. In the latter condition, adults performed worse on reproducing the noun–particle pairings for the two consistent nouns, compared with adults in the lexically consistent condition. Children, on the other hand, performed equally well in these two conditions. This suggests that adult learners prefer variation (represented by the two meaningless particles), to be lexically conditioned. This is compatible with the preference we found for complex > random grammar. The complex grammar is inconsistent overall with respect to the ordering of affix categories, but this variation is conditioned on the level of individual affixes. The random grammar has variable ordering without conditioning on any level. Thus, our results suggest a preference for variation to be conditioned by specific affixes, just as Wonnacott et al. (2017) results show an (adult) preference for variation to be conditioned by specific lexemes.

5.3. The link between learning biases and cross-linguistic distributions

A series of questions remain open as to how a category clustering bias in individual learners may affect language structure, given that individual biases are mediated by the social structure of language learning. Our experiment demonstrates only an individual learning bias in adults, without having explored its interaction with the pressures at play during transmission or communicative interaction. However, natural languages evolve through a repeated cycle of learning and use, that is, over many generations of children and adults learning languages and reproducing them with modifications, which are later transmitted. To understand how individual learning biases influence natural languages, we need to understand how those biases operate in the population-level convergence on communicative conventions.

Blythe and Croft (2021) provide a recent model of how individual speakers produce population-level language change. They treat the situation of a single meaning with multiple competing forms for expressing that meaning, and model this as a probability distribution over the competing forms, $P(M) = \sum (p(f_1) \ldots p(f_n))$. Each individual at a given point in time has their own particular probability distribution, and when individuals interact, their probability distributions become more similar to one another. The model thus provides a mechanism by which individual biases can spread through a population, via linguistic interaction with accommodation. To investigate this as a model of the category clustering bias, we would first require a different experimental design that measures how an individual participant responds to the choice between clustered and unclustered grammars. This would provide data on individuals' selection among competing forms as opposed to the current experiment that estimates differing learning outcomes between individuals in different conditions. We would also require data on how individuals with different morphological grammars accommodate when they interact, following a model of language change such as Blythe and Croft (2021). Given such data, we might then investigate whether individual biases, modulated by interaction and accommodation (see Wade & Roberts, 2020), could plausibly produce a natural language distribution of the type observed in Mansfield et al. (2020).

As outlined earlier, iterated learning experiments explore individual versus group dynamics in human participants, modeling the cultural evolution of linguistic systems via transmission chains with multiple generations of learners/users in the lab. Each generation of participants (one or more) is taught a miniature language and is later asked to reproduce it during testing; the output is then passed on to the next generation. These studies show that the structure of iterative transmission has an important mediating effect on individual learning biases over time (Griffiths & Kalish, 2007; Kirby et al., 2008; Reali & Griffiths ,2009; Smith & Wonnacott, 2010; Smith et al., 2017). On the one hand, weak biases may be amplified by multiple generations that incrementally shift grammatical structure in the same direction. This could mean that the weak bias we found for simple grammars would lead to a clear typological bias as morphological structures grammaticalize over many generations. On the other hand, even strong biases in individuals may be neutralized if each generation of learners acquires their input from a mixture of speakers with different biases (Smith et al., 2017). This could permit the development and persistence of complex and even random grammars in some languages, even in the face of individual biases against these grammars. The size of speech communities

may also play a role, as larger networks of interlocutors have been found to produce more systematic morphology than smaller networks (Raviv, Meyer, & Lev-Ari, 2019). To draw further conclusions about how a learning bias toward category clustering might relate to typological distributions, one could design iterated learning experiments that explore how groups of different sizes, composed of individuals with different clustering biases, gradually converge on simple, complex (or perhaps even random) grammars.

Further unexplored dimensions lie in potential differences between child and adult learners, visual learning of text versus auditory stimuli, and differences based on pre-existing language competence. Our experiment shows a bias in textual learning among second-language adult learners who are speakers of English. But it remains to be seen whether this bias is also found in auditory learning, whether it is shared by children and by adult speakers of languages with unclustered morphology. We noted above that learners of the simple grammar, who achieved overall more accurate learning, may have used a more compositional strategy than their less successful counterparts in the complex grammar. But if children begin learning a language with more holistic memorization, and only gradually develop compositional strategies (Bannard & Matthews, 2008; Lieven, Pine, & Baldwin, 1997; Peters, 1977; Tomasello, 2003; Wray, 2002), then this could reduce the impact of a bias toward grammars with category clustering.

Any differences in the clustering bias among types of learners could also be expected to have typological consequences: If only learners of a second language exhibit the category clustering bias, then the typological bias should be limited to languages with substantial numbers of second-language learners and show geographical patterns of areal spread through language contact (Lupyan & Dale, 2010; Nichols, 2003; Trudgill, 2011; Widmer, Jenny, Behr, & Bickel, 2020). But if the learning bias also applies to language-internal variation and change, that is, when learning new variants of one's own language, then the typological bias should be general and not limited to the results of language contact. The results of Mansfield et al. (2020) did not reveal obvious geographical patterns, tentatively suggesting a more general learning bias. But more data are needed to assess this while controlling for potential effects of community size (as noted above).

The English-language experience of our participants may also have influenced the results, and if so, this will only become clear if further experiments are conducted with speakers of other languages. English has fairly limited inflectional morphology, and the only clear instance of paradigmatic affix alternation involves verbal suffixes (*-s, -ed, -ing*). These affixes conform to category clustering only to a limited extent since they do not instantiate strictly the same category (*-s* expresses tense, person and number; *-ed* tense only, and *-ing* aspect only). It is possible that speakers of languages with more extensive affix clustering (e.g., Swahili), complex-type affixation (e.g., Arabic), or random-type affixation (e.g., Chintang) may be better at learning these respective morphological grammars relative to English speakers. On the other hand, as we showed above, category clustering does appear to be a very general property of natural languages, spanning both morphology and syntax. We might therefore expect that the clustering bias would be found to some degree in speakers of all languages.

6. Conclusion

Affixes expressing the same grammatical category tend to appear in the same morphological position in the word. While this "category clustering" effect is assumed by most theories of morphology and has previously been demonstrated in the typological outcomes of language evolution, this study has been the first to demonstrate that it is also a bias in individual learning. Hypothesizing that the typological bias toward clustering is at least partially caused by a learning bias, we taught adult English speakers a miniature artificial language with experimental conditions that did or did not conform to category clustering. The language consisted of stems expressing shape and pairs of suffixes expressing color and number. In line with our hypothesis, we found that a version of the language conforming to category clustering was learned more accurately than two different versions that violate category clustering. In one non-clustered version (complex), each specific suffix combination has a fixed order, but this order is not consistent for different combinations of color and number suffixes. In the other non-clustered version (random), there is no fixed ordering of suffixes, and they are randomly ordered on each presentation. We found that of these two non-clustered grammars, learners performed better with the complex grammar than the random grammar. Although we did not have any specific hypothesis regarding complex versus random grammars, these findings are consistent with cross-linguistic surveys, where many languages show some degree of complex affix ordering, but random ordering is rare.

Although the experimental results supported our hypothesis of a learning bias toward category clustering, there remain several open questions about how this individual bias may be associated with the evolution of affix ordering in natural languages. Language transmission within and across generations of learners has an important mediating effect, which may amplify or mask individual biases depending on the strength of each of the pressures at play during transmission. Furthermore, children are known to have different learning biases from adults, and previous language experience may influence the learning of new linguistic input. Our results also suggested potential differences in compositional versus whole-word learning strategies, which may lead to different results in children, compared to adults. In our experiment, we have demonstrated the category clustering bias only for adults, who were already speakers of English. While further research is still required, the results of the current study support the hypothesis that the dominance of category clustering in inflectional morphology is at least partially produced by a learning bias.

Acknowledgments

We would like to thank all the participants who took part in this research. This research was supported by the Australian Research Council awards DE180100872 and DP180103600, the ARC Centre of Excellence for the Dynamics of Language CE140100041, and the NCCR Evolving Language, Swiss NSF Agreement Nr. 51NF40_180888.

Open access publishing facilitated by The University of Melbourne, as part of the Wiley-The University of Melbourne agreement via the Council of Australian University Librarians.

WOA Institution: The University of Melbourne Blended
DEAL: CAUL 2022

Notes

1 On the other hand, Swahili negative prefixes violate category clustering, with distinct "general negative" and "negative subjunctive" prefixes appearing in different positions (Stump, 1997, p. 221).
2 It is sometimes proposed that category clustering in morphology is a historical residue of category clustering in syntax. Givón (1971) proposes that linear ordering in morphology reflects the linear ordering of earlier phrase structures, which are the historical sources of morphology. However, it remains unclear to what extent this is the case, given that morphological ordering can also deviate substantially from phrasal sources (Anderson, 1980). Also, even where the historical scenario is genuine, it would merely shift the clustering bias from morphology to syntax. As such, it would still require explanation.
3 The pre-registered design and analysis plan is accessible at https://aspredicted.org/rz96j.pdf. Note that the analysis plan does not conform to the analysis presented here in several ways (see the Results section and Footnote 9).
4 Informed consent was obtained from all participants, following a research protocol approved by the Human Ethics Advisory Group at the University of Melbourne, HREC #1852093.1.
5 A pilot of this experiment ($N = 20$) was also run before data collection in order to test the software. These data are not included in our analysis.
6 Color is sometimes said to be unattested as a grammatical category (e.g., Seifart, 2010, p. 726), although at least color versus non-color distinctions do occasionally show up in inflection (Bickel, 2017).
7 We adopted this set of shapes from experiments testing for clairvoyance. In the original experiments, the images were printed on "Zener cards," visible only to one participant, while another was invited to use clairvoyance to identify the image (Laycock, 1989). Some readers may be familiar with this experimental procedure from the opening scene of the film *Ghostbusters* (Reitman, 1984).
8 We also screened participants on their ability to recall the instructions and to distinguish the colors utilized in the experiment; they could not participate if they could not adequately summarize the instructions provided and/or distinguish blue, red, and green.
9 In the pre-registration, we proposed to analyze results only in terms of seen/unseen training items, but after running the experiment, we realized that more nuance and power could be added by quantifying the specific number of views.
10 A beta-binomial allows us to account for the over-dispersion we expect in learning tasks, where variability decreases as a function of time if indeed learners eventually learn. The model estimates the distribution of probabilities of success across cases; each binomial count observation has its own probability of success (McElreath, 2016, p. 346).

11 https://osf.io/mwj54/

12 We report 95% uncertainty intervals throughout the Results section because effective sample sizes are >10k across fixed effects except for the number of training views which is 9221, suggesting highly reliable posterior estimates (Kruschke, 2014, p. 183).

References

Anderson, S. R. (1980). On the development of morphology from syntax. In J. Fisiak (Ed.), *Historical morphology* (pp. 51–70). The Hague: Mouton.

Anderson, S. R. (1992). *A-morphous morphology*. Cambridge, UK: Cambridge University Press.

Arnott, D. W. (1970). *The nominal and verbal systems of Fula*. Oxford, England: Oxford University Press.

Baerman, M., & Corbett, G. (2012). Stem alternations and multiple exponence. *Word Structure*, *5*(1), 52–68.

Baker, M. (1985). The mirror principle and morphosyntactic explanation. *Linguistic Inquiry*, *16*(3), 373–415.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176.

Bickel, B. (2017). Belhare. In G. Thurgood, & R. J. La Polla (Eds.), *The Sino-Tibetan languages* (2nd ed., pp. 546–570). London: Routledge.

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, I. P., Rai, M., Rai, N. K., & Stoll, S. (2007). Free prefix ordering in Chintang. *Language*, *83*(1), 43–73.

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K. A., Rießler, M., Bierkandt, L., Zúñiga, F., & Lowe, J. B., 2017. *The AUTOTYP typological databases*. Retrieved from https://github.com/autotyp/autotyp-data/tree/0.1.0

Blythe, R. A., & Croft, W. (2021). How individuals change language. *PLOS One*, *16*(6), e0252582.

Bouchard, D. (1998). The distribution and interpretation of adjectives in French: A consequence of Bare Phrase Structure. *De Gruyter Mouton*, *10*(2), 139–184.

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

Bürkner, P. -C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam, The Netherlads: John Benjamins.

Bybee, J. L., Perkins, R. D., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago, IL: Chicago University Press.

Crysmann, B. & Bonami, O. (2016). Variable morphotactics in information-based morphology. *Journal of Linguistics*, *52*(2), 311–374.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Culbertson, J., Franck, J., Braquet, G., Barrera Navarro, M., & Arnon, I. (2020). A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, *204*, 104392.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., & Newport, E. L. (2017). Innovation of word order harmony across development. *Open Mind*, *1*(2), 91–100.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Dryer, M. S. (2013a). Order of subject, object and verb. In M. S. Dryer, & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81

Dryer, M. S. (2013b). Prefixing vs. suffixing in inflectional morphology. In M. S. Dryer, & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/26

Embick, D., & Noyer, R. (2001). Movement operations after syntax. *Linguistic Inquiry*, *32*(4), 555–595.

Fedzechkina, M., Jaeger, T. F., & Newport, Elissa L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Foley, W. A., & Van Valin, R. D. (1984). *Functional syntax and universal grammar*. Cambridge, UK: Cambridge University Press.

Givón, T. (1971). Historical syntax and synchronic morphology: An archaeologist's field trip. *Chicago Linguistic Society*, *7*(1), 394–415.

Gleason, H. A. Jr. (1955). *Workbook in descriptive linguistics*. New York: Holt, Rinehart & Winston.

Good, J. (2016). *The linguistic typology of templates*. Cambridge, UK: Cambridge University Press.

Greenberg, J. (1963). *Universals of language*. London: MIT Press.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale, & S. J. Keyser (Eds.), *The view from building* (Vol. *20*, pp. 111–176). Cambridge, MA: MIT Press.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

Himmelmann, N. P. (2014). Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language*, *90*(4), 927–960.

Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, *24*(6), 876–909.

Jackendoff, R., & Audring, J. (2019). *The texture of the lexicon: Relational Morphology and the Parallel Architecture*. Oxford, England: Oxford University Press.

Keresztes, L. (1998). Mansi. In D. Abondolo (Ed.), *The Uralic languages* (pp. 387–427). London: Routledge.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York: Academic Press.

Laycock, D. C. (1989). *Skeptical, a handbook on pseudoscience and the paranormal*. Canberra: Canberra Skeptics.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structue. *PLoS ONE*, *5*(1), 1–10.

Maldonado, M., Saldana, C. & Culbertson, J. (2020). Learning biases in the relative order of person and number markers. Proceedings of the 50th Annual Meeting of the North East Linguistic Society (NELS 50). Amherst, MA. 163-176

Manova, S. (2015). *Affix ordering across languages and frameworks*. Oxford, England: Oxford University Press.

Manova, S., & Aronoff, M. (2010). Modeling affix order. *Morphology*, *20*(1), 109–131.

Mansfield, J. (2019). *Murrinhpatha morphology and phonology*. Berlin: De Gruyter Mouton.

Mansfield, J., Stoll, S., & Bickel, B. (2020). Category clustering: A probabilistic bias in the morphology of argument marking. *Language*, *96*(2), 255–293.

Marslen-Wilson, W. D. (2007). Morphological processes in language comprehension. In G. Ramchand, & C. Reiss (Eds.), *The Oxford handbook of psycholinguistics* (pp. 175–193). Oxford, England: Oxford University Press.

Martin, A., & Culbertson, J. (2020). Revisiting the suffixing preference: Native-language affixation patterns influence perception of sequences. *Psychological Science*, *31*(9), 0956797620931108.

Martin, A., Ratitamkul, T., Abels, K., Adger, D., & Culbertson, J. (2019). Cross-linguistic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*, *5*(1).

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. London: CRC Press.

Mellis, A. M., & Bickel, W. K. (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction*, *115*(10), 1960–1968.

Nichols, J. (2003). Diversity and stability in language. In B. Joseph, & R. D. Janda (Eds.), *The handbook of historical linguistics* (pp. 283–310). Oxford, England: Blackwell.

Peters, A. M. (1977). Language learning strategies: Does the whole equal the sum of the parts? *Language*, *53*(3), 560–573.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Raviv, L. Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262.

Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. *Cognition*, *210*, 104620.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Reitman, I. (1984). *Ghostbusters*. Culver City, CA: Columbia Pictures.

Saldana, C., Kirby, S., Truswell, R., & Smith, K. (2019). Compositional hierarchical structure evolves through cultural transmission: An experimental study. *Journal of Language Evolution*, *4*(2), 83–107.

Saldana, C., Oseki, Y., & Culbertson, J. (2021). Cross-linguistic patterns of morpheme order reflect cognitive biases: An experimental study of case and number morphology. *Journal of Memory and Language*, *118*, 104204.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.

Seifart, F. (2010). Nominal classification. *Language and Linguistics Compass*, *4*(8), 719–736.

Smith, K., Perfors, A. Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

St Clair, M.C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Stan Development Team. 2020. *Stan modeling language users guide and reference manual* (version 2.21.2). Retrieved from http://mc-stan.org/

Stump, G. T. (1993). Position classes and morphological theory. In G. Booij, & J. van Marle (Eds.), *Yearbook of morphology 1992* (pp. 129–180). Amsterdam: Kluwer Academic Publishers.

Stump, G. T. (1997). Template moprhology and inflectional morphology. In G. Booij, & J. van Marle (Eds.), *Yearbook of morphology* (pp. 217–241). Amsterdam: Kluwer Academic Publishers.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure*. Cambridge, UK: Cambridge University Press.

Stump, G. T. (2015). *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge, UK: Cambridge University Press.

Tabullo, Á., Arismendi, M., Wainselboim, A., Primero, G. Vernis, S., Segura, E., Zanutto, S, & Yorio, A. (2012). On the learnability of frequent and infrequent word orders: An artificial language learning study. *Quarterly Journal of Experimental Psychology*, *65*(9), 1848–1863.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford, England: Oxford University Press.

Wade, L., & Roberts, G. (2020). Linguistic convergence to observed versus expected behavior in an alien-language map task. *Cognitive Science*, *44*(4), e12829.

Waugh, L. R. (1977). *A semantic analysis of word order: Position of the adjective in French*. Leiden: Brill.

Wendorf, C. A. (2004). Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics*, *3*(1), 47–57.

Widmer, M., Mathias, J., Wolfgang, B., & Bickel, B. (2020). Morphological structure can escape reduction effects from mass admixture of second language speakers: Evidence from Sino-Tibetan. *Studies in Language*, *45*(4), 707–752.

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, *95*, 36–48.

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.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material