Experimental evidence for the influence of structure and meaning on linear order in the noun phrase

Alexander Martin
Centre for Language Evolution, University of Edinburgh
3 Charles St, Edinburgh EH8 9AD, UK
alxndr.martin@gmail.com

Annie Holtz
Centre for Language Evolution, University of Edinburgh
3 Charles St, Edinburgh EH8 9AD, UK
annie.holtz@sms.ed.ac.uk

Klaus Abels
Division of Psychology and Language Sciences, University College London
Chandler House, 2 Wakefield Street, London WC1N 1PF, UK
k.abels@ucl.ac.uk

David Adger
School of Language, Linguistics and Film, Queen Mary University of London
Mile End Road, London E1 4NS, UK
d.j.adger@qmul.ac.uk

Jennifer Culbertson
Centre for Language Evolution, University of Edinburgh
3 Charles St, Edinburgh EH8 9AD, UK
jennifer.culbertson@ed.ac.uk

Abstract Recent work has used artificial language experiments to argue that hierarchical representations drive learners' expectations about word order in complex noun phrases like *these two green cars* (Culbertson & Adger 2014; Martin, Ratitamkul, et al. 2019). When trained on a novel language in which individual modifiers come after the Noun, English speakers overwhelmingly assume that multiple nominal modifiers should be ordered such that Adjectives come closest to the Noun, then Numerals, then Demonstratives (i.e., N-Adj-Num-Dem or some subset thereof). This order transparently reflects a constituent structure in which Adjectives combine with Nouns to the exclusion of Numerals and Demonstratives, and Numerals combine with Noun+Adjective units to the exclusion of Demonstratives. This structure has been also claimed to derive frequency asymmetries in complex noun phrase order across languages (e.g., Cinque 2005). However, we show that features of the methodology used in these experiments potentially encourage participants to use a particular metalinguistic strategy that could yield this outcome without implicating constituency structure. Here, we use a more naturalistic artificial language learning task to investigate whether the preference for hierarchy-respecting orders is still found when participants do not use this strategy. We find that the preference still holds, and, moreover, as Culbertson & Adger (2014) speculate, that its strength reflects structural distance between modifiers. It is strongest when ordering Adjectives relative to Demonstratives, and weaker when ordering Numerals relative to Adjectives or

Demonstratives relative to Numerals. Our results provide the strongest evidence yet for the psychological influence of hierarchical structure on word order preferences during learning.

Keywords: learning bias; syntax; typology; artificial language learning

1 Introduction

It has long been argued that hierarchical structure is a key feature of human language (e.g., Chomsky 1965; Gleitman & Newport 1995). Hierarchical representations form the basis of many formal theories of syntax, encoding constituency relations among the linguistic sub-units that comprise phrases and sentences, and connecting directly to theories of semantic interpretation (Adger 2003; Heim & Kratzer 1998). However, there is ample controversy over the degree to which this kind of hidden structure underlies human linguistic competence, and whether it plays a major role in language comprehension and production. In alternative theories, surface features like linear order and in particular frequent linear sequences of words, are the primary representational units (e.g., Goldberg 2006; Tomasello 2003; McCauley & Christiansen 2017; Frank & Bod 2011). In recent work, artificial language learning experiments have been used to argue for the importance of hierarchical structure, rather than linear order, in driving learners' early inferences about complex noun phrase word order (Culbertson & Adger 2014; Martin, Ratitamkul, et al. 2019). It was shown that when the input under-determines the word order of a language, learners generalise in a way that conforms to a hierarchical representation of the noun phrase rather than the linear order of their native language. Those experiments, however, suffer from a potentially serious methodological issue. Here, we turn a critical eye on those previous results, resolving a problematic feature of the design of those experiments, and providing new evidence for the relationship between hierarchical structure, meaning, and linear order in the noun phrase.

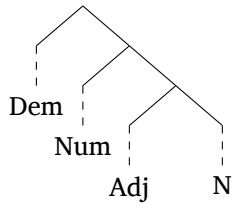
1.1 Hierarchy in the noun phrase

In a complex noun phrase like *these two green cars*, a Noun (N) is modified by an Adjective (Adj), a Numeral (Num), and a Demonstrative (Dem). The surface linear order (in English) is Dem-Num-Adj-N, where three modifiers give us additional detail we can use to pick out exactly which cars are under discussion. However, many syntactic and semantic theories of the noun phrase posit layers of structure that are not obvious from the linear order.

In particular, it has been argued that Adjectives form a syntactic constituent with the Noun to the exclusion of Num and Dem. One source of evidence for this comes from *constituency tests*–a standard tool in theoretical syntax for detecting hierarchical structure (e.g., Adger 2003; Abels 2015). These tests show that Noun + Adj behaves as a linguistic unit. For example, in "I like these green cars and Ally likes those." the unit "green cars" is understood in the second conjunct even though it is not phonetically present. Contrast

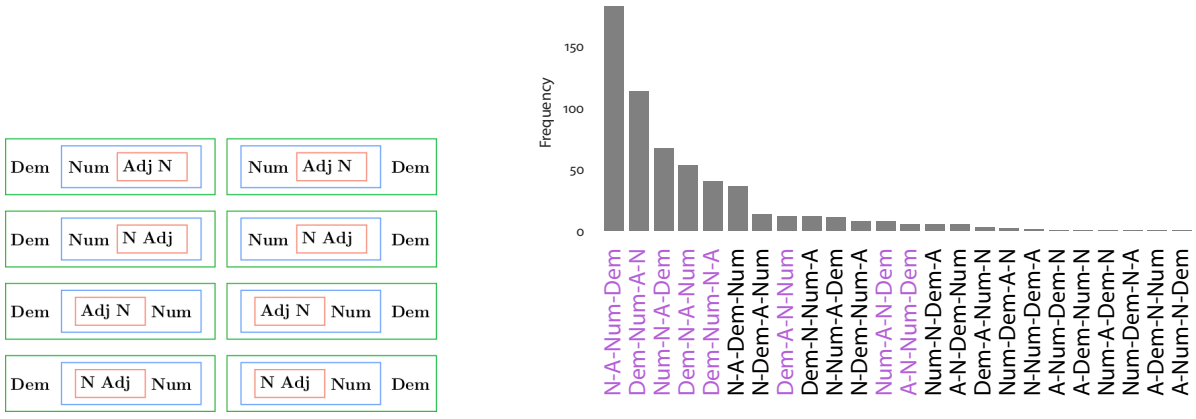
this with "I like these green cars and Ally likes bikes", where this cannot be understood, analogously, as "Ally likes these green bikes". The explanation for this is that there is a deletion process, allowing speakers to omit certain words in cases of semantic parallelism, under the condition that these words must form a constituent. Multiple distinct diagnostic tests like these converge on a structure where not only are the Adjective and Noun a constituent to the exclusion to the Demonstrative, but also where the Numeral forms a constituent with the Adjective-Noun unit excluding the Demonstrative. For the noun phrase, the constituency structure is shown in (1).

(1)



The exact source of this hierarchical structure is debated, but many approaches assume it is derived from semantic composition (Abels & Neeleman 2012; Ernst 2002; Steedman 2018) or conceptual structure (Rijkhoff 1990; 2004; Hurford 1987; Culbertson, Schouwstra & Kirby to appear). For example, in formal semantics, it is typically argued that Adjectives and Nouns are of the same semantic type, and that when they combine, the result is a nominal predicate (e.g., Partee 1987). By contrast, Numerals have been analysed as the values of measure functions that map nominal predicates to countable units (e.g., Krifka 1995). Empirical evidence for this approach is the fact that, while languages commonly require classifiers for counting using numerals, they do not require classifiers for modification by adjectives. In other words, Numerals are a distinct semantic entity from Adjectives, and compose compositionally with a unit that can include either a Noun or an Adjective together with a Noun. Demonstratives are distinct from both Adjectives and Numerals, mapping nominal predicates to individuals via the speech situation they are used in (e.g., Elbourne 2008). Like other elements, such as definite articles, Demonstratives thus semantically compose with the Noun only after composition with Adjectives and Numerals. Composition order is sometimes referred to in terms of scope: Demonstratives take highest scope, then Numerals, then Adjectives.

A good deal of work in semantics has focussed on pinning down exactly how elements like Demonstratives, Numerals, and Adjectives differ from one another in terms of their meaning. However, from a syntactic perspective, it is sufficient to assume that the semantic distinctions are linked to syntactic categories which are hierarchically organised as in (1). This hierarchy can be used to underpin an account of noun phrase *word order* variation across languages. This is because it has the potential to explain a striking asymmetry in the frequency of word order patterns across the world's languages. The tree in (1) is an abstract representation which does not encode linear order. To linearise it, one must make a binary decision at each branch, [Adj N] or [N Adj], then [Num [Adj N]] or [[Adj N] Num] and so on. There are eight distinct linear orders that result from this (shown in fig. 1(a)), and these have a particular feature in common: they all have Adj closest to the Noun, then Num, and then Dem farthest away. Other logically possible



(a) The eight homomorphic orders.

(b) Typological distribution of homomorphic orders (highlighted in colour) among all 24 possible orders (Dryer 2018).

Figure 1: Homomorphic orders and their typological distribution..

linear orders—there are 16 of them—cannot be the direct result of linearising (1) in this fashion; for example, N-Dem-Num-Adj has both Dem and Num intervening between N and Adj, and is not derivable from linearising each choice point.[1]

If one counts up the unmarked orders for a reasonably large sample of languages, the eight orders in fig. 1(a) are consistently among the most common (e.g., Greenberg 1963; Cinque 2005; Dryer 2018). For example, nearly half of the world's languages have the order N-Adj-Num-Dem (like Thai) or Dem-Num-Adj-N (like English). Orders like Dem-Num-N-Adj (e.g., French), Num-N-Adj-Dem (e.g., Basque) and Dem-N-Adj-Num (e.g., Burmese) are also commonly attested. By contrast, other possible orders are systematically less frequent and in some cases (e.g., Adj-Num-Dem-N) have never been found, at least not as the neutral (or default) order, in a known language. The estimated frequencies from Dryer (2018) are shown in fig. 1(b).

This apparent typological asymmetry is potentially a strong additional source of evidence for the hierarchical representation of the noun phrase outlined above. In particular, if we assume that all humans represent the underlying structure of the noun phrase as in (1), we can explain why these eight orders are more common than the others; if humans are biased in favour of mapping structure to linear order in a transparent way, then these orders should be preferred. Previous work has referred to this as a bias for isomorphism

[1] Of course, one could argue that these orders are derived by syntactic movement. Languages allow alternative orders in the noun phrase to achieve effects of emphasis. For example, in French, the Adjective "énorme" typically follows the Noun (e.g., "une aire énorme"–*an enormous area*), but can be moved prenominally for emphasis (e.g., "c'était un énorme problème pour lui". It could be that some languages have *neutral* orders that are derived via movement. This is exactly what is argued for by Cinque (2005) and Abels & Neeleman (2012).

(Culbertson & Adger 2014), though homomorphism is in fact a more accurate term.[2] Indeed, previous accounts of noun phrase word order consider (1) as a kind of default (Cinque 2005; Abels & Neeleman 2012; Steddy & Samek-Lodovici 2011; Dryer 2018).

However, the fact that many more than the eight homomorphic orders orders of N, Adj, Num, and Dem are attested in some language (Dryer 2018: though not all of these orders are necessarily unmarked), combined with the lack of work that the constituency relations posited above in fact hold universally (i.e., even in non-homomorphic languages), calls into question whether the hierarchical representation hypothesised for the noun phrase is real. An alternative possibility is that the power law distribution in fig. 1(b) is not the result of a bias, but of historical accident (Piantadosi & Gibson 2014). Learners could instead either represent noun phrases purely in terms of their linear order, or could infer an underlying representation which is derived from linear order, rather than the reverse.

1.2 Experimental evidence for the role of hierarchy in learning

Recently, artificial language learning experiments have been used to test the hypothesis that noun phrase order is influenced by a homomorphism bias (Culbertson & Adger 2014; Martin, Ratitamkul, et al. 2019). In particular, these experiments tested whether learners use hierarchical structure rather than linear order when required to generalise beyond their input in a new language. The prediction was that if participants represent the noun phrase in terms of a structure like that in (1), they should automatically assume a homomorphic order. In Culbertson & Adger (2014), English-speaking participants were trained on a new version of English, where all nominal modifiers appeared *after* the Noun. Thus, phrases like *green car* were shown as *car green*. Participants were then asked to guess how noun phrases with two modifiers (e.g., *these green cars*) might be produced in the new version of English. Overwhelmingly, and regardless of the combination of modifiers they were trained on,[3] participants chose homomorphic orders (e.g., *cars green these*) over non-homomorphic ones (e.g., *cars these green*). This is striking, since they had no input to indicate that a homomorphic order should be preferred over a non-homomorphic one in this new language. The authors further showed that if learners were basing their guesses on the bigram statistics of English (i.e., if their representation of English linear order determined their responses) they should choose non-homomorphic orders over homomorphic ones. For example, a participant might not have a preference for one order of modifiers over the other, but if they know that in English the word *this* precedes the word *green* in complex noun phrases, they should choose the order *car this green*. Martin,

[2] Strictly speaking, an isomorphism is defined by the ability to reverse the mapping and obtain the same structure. While this is true of orders like Thai (N-Adj-Num-Dem) and English (Dem-Num-Adj-N) (where the exact hierarchical relations amongst elements are maintained), a linearisation like that of French (canonically Dem-Num-N-Adj) would be more appropriately deemed *homo*morphic, since reversing the French mapping does not yield the exact underlying structure; in that structure, Numerals are further from the Noun than Adjectives, but both are equally far from the Noun in French. Crucially, though, the hierarchical relations are never reversed in homomorphic linearisations. Because we hypothesise that the latter is the crucial characteristic, we switch here to the term homomorphic.

[3] In the first experiment reported in that paper, participants saw noun phrases with Demonstratives or Adjectives, Numerals or Adjectives, *or* Demonstratives or Numerals and were then tested on two-modifier phrases that combined the two types of modifiers they were trained on.

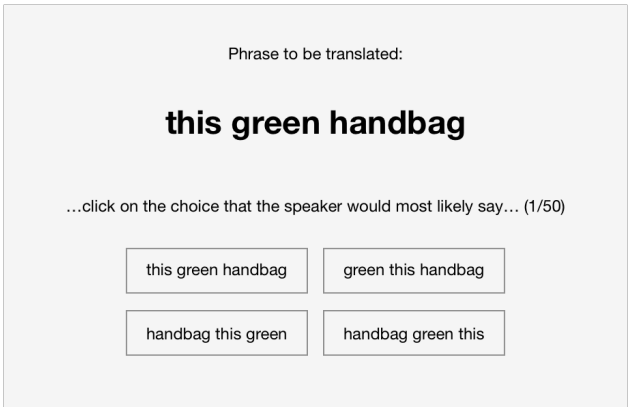


Figure 2: Example of a two-modifier test trial in Culbertson & Adger (2014).

Ratitamkul, et al. (2019) replicated this effect with Thai speakers, whose native language order is N-Adj-Num-Dem (using versions of the materials translated into Thai).

Culbertson & Adger (2014) and Martin, Ratitamkul, et al. (2019) concluded that speakers must be relying on abstract hierarchical representations like (1) to determine word order in the experiment. There are two interpretations consistent with this result. It could reflect Thai and English speakers' prior knowledge of a homomorphic language. For example, if hierarchical relations among elements in the noun phrase are derived from experience with linear order in speakers' native language, then both English and Thai speakers will have inferred a structure like (1). They could then use this structure to determine the likely relative order of modifiers in a new language. The previous results are also consistent with the hypothesis that the underlying structure in (1) is universal, and therefore is used to infer word order in a new language regardless of speakers' prior experience with a specific language. However, there is a methodological problem with these previous experiments which impacts on the robustness of such interpretations. In both Culbertson & Adger (2014) and Martin, Ratitamkul, et al. (2019), participants were presented with a written "phrase to be translated" and a set of written response options. In training trials, single-modifier noun phrases were backwards versions of their native language. If participants noticed this, they could have adopted a strategy of "flipping" the native language phrase to determine their choice. In testing trials (such as fig. 2), this flipping strategy would result in homomorphic order without any representation of the hierarchical structure.

In Culbertson & Adger (2014), this possibility was addressed by pointing to a qualitative difference in results across combination types: the homomorphism preference appeared to be strongest when ordering Dem-Adj-N combinations (compared to Dem-Num-N or Num-Adj-N combinations). The authors argued that this would make sense if participants' responses are driven by the underlying constituency structure, as these two modifier types are structurally more distant from one another than Numeral and Demonstrative, or Adjective and Numeral, see (1). By contrast, this difference is not predicted by a flipping strategy. However, previously unanalysed self-reported response strategies from those studies (Experiment 1 of Culbertson & Adger (2014) and Experiment 2 of Mar-

Table 1: Counts of participants reporting different strategies in Experiment 1 of Culbertson & Adger (2014) (N=87) and in the Thai-language equivalent reported by Martin, Ratitamkul, et al. (2019) (N=112).

	Culbertson & Adger	Martin et al.
Flipping	40	42
Noun flip	7	11
(Non-)Homomorphic	4	14
L2	0	18
No strategy or no response	26	21
Other	10	6

tin, Ratitamkul, et al. (2019)) in fact suggest the methodological problem is still a serious one. As shown in table 1, flipping is by far the most common strategy reported.[4]

These self-reports, together with the highly artificial nature of the experimental design in these previous experiments, point to the need for a conceptual replication. Below, we report three experiments, run online and in a controlled laboratory environment, which use a more naturalistic artificial language learning task. The experimental results still point to sensitivity to hierarchical structure in the form of a homomorphism preference, although as we discuss below, this turns out to depend on the nature of the task. Importantly, no participants report using any flipping-based strategy to give their responses. We also confirm, across these experiments, the effect of modifier-type on the strength of the homomorphism preference (i.e., a consistent difference between conditions). We conclude the paper by discussing what differences in homomorphism across modifier types might tell us about the interaction between meaning, structure and linear order in this domain.

2 Experiment 1

Following Culbertson & Adger (2014), we used an artificial language learning paradigm in which participants are taught phrases with a Noun and a single modifier (either an Adjective and a Demonstrative, a Numeral and a Demonstrative, or an Adjective and a Numeral), and must guess the relative order of modifiers when both meanings are present. For example, participants may learn how to say *red ball* and *two balls* and then must infer

[4] Any strategy that used words such as "reverse", "flip", or "backwards" was coded as *flipping* (e.g., "I decided to read everything backwards"). We coded responses that only mentioned moving the position of the Noun as *Noun flip* (e.g., "I moved the noun to the beginning"). While this is also a metalinguistic strategy, it does not yield homomorphic order preferences. Strategies that described the specific order given were coded as (non-)homomorphic (e.g., a Thai non-homomorphic response strategy report was "I put colour words first, then number words, then nouns"). Strategies that compared responses to other languages (e.g., Thai participants who said they responded using English order) were coded as *L2*. Note that no participants explicitly referred to their native language order. Some participants reported not using any specific strategy at all, gave incomprehensible responses (e.g., "I visualised the noun in my mind, then pointed to it/them"), or did not respond to the question.

how to say *two red balls*. Crucially, in contrast with previous work, we used completely novel stimuli and did not present written native language equivalents of the phrases participants were learning. This was done to reduce the possibility, present in Culbertson & Adger (2014) and Martin, Ratitamkul, et al. (2019), that participants would visually "flip" their native language word order to translate into the artificial language they were learning. Participants each learned two kinds of modifiers, depending on the condition they were assigned to.

2.1 Methods

2.1.1 Stimuli

The artificial language had seven lexical items. There were three Nouns meaning *feather*, *ball*, and *mug*, visually presented as orthographic <éyày>, pronounced /éjè/, <úhù>, pronounced /úhù/, and <ítì>, pronounced /ítì/, respectively. The modifier meanings were two Adjectives (*red* and *black*), two Numerals (*two* and *three*), and two Demonstratives (proximal and distal). Labels for these modifier classes were created in pairs: <pùkú> pronounced /púkù/, <tákà> pronounced /tàká/ and <hìmí> pronounced /hìmí/, <hònó> pronounced /hónò/. We privileged within-pair similarity (so /púkù/ and /tàká/ both contain only voiceless stops for example) to facilitate the learning process. The two pairs of stimuli were assigned meanings according to the condition participants were in: in the Dem-Adj condition, the fricative-initial words were Demonstratives and the stop-initial words were Adjectives; in the Dem-Num condition, the fricative-initial words were Demonstratives and the stop-initial words were Numerals; in the Num-Adj condition the fricative-initial words were Numerals and the stop-initial words were Adjectives. Stimuli were produced by a trained phonetician. All stops were produced with near zero VOT and each syllable was produced with either a high or a low tone.[5]

As is standard in artificial language learning experiments with novel lexical items, participants had to infer the meanings (and linguistic categories) of these items from visual stimuli. We used pictures of simple cartoon scenes. Objects were depicted on a table behind which stood a cartoon girl. In trials featuring the Noun alone, or the Noun with an Adjective and/or Numeral, the girl was simply shown behind the table. In trials featuring a Demonstrative, the girl was shown pointing to an object or objects (either near to her, or on the other side of the table from her). The presence of the girl and table on all trials was meant to keep Demonstrative trials from being more visually salient (or complex). When no adjectival meaning was expressed, objects were drawn in light grey; objects were only coloured in (in red or black) on trials involving Adjectives. Examples of the visual stimuli for single modifier trials are shown in fig. 3. Participants were given

[5] We designed the language to encourage participants to perceive it as a real "foreign" language. Therefore, while the words do not overtly contradict English phonotactics, they are not particularly English-like. This makes them difficult to learn. Piloting suggested that keeping the vocabulary relatively small would be necessary. Note that the orthographic accents were only present to make the items appear less English-like, and did not necessarily correspond to the tone the items were pronounced with. We manipulated tone on the stimuli as we plan to use the same experiment design in other linguistic populations in the future (including Thai, which has lexical tone).

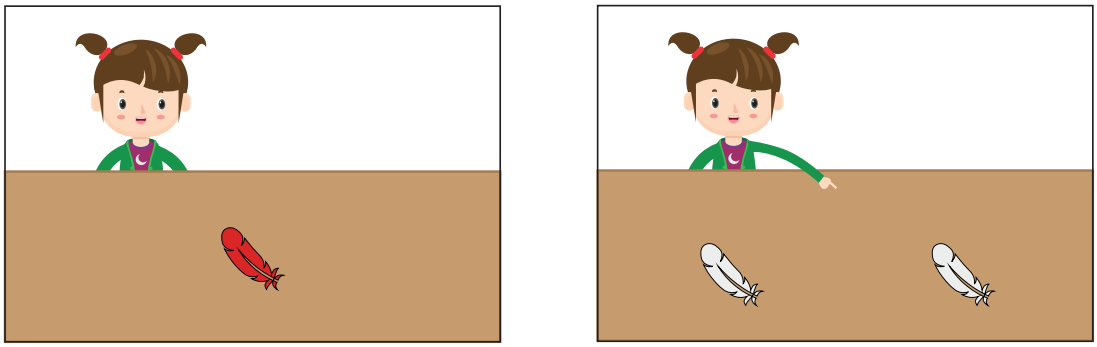


Figure 3: Single modifier trial visual stimuli examples. On the left, an example of an Adjective trial, meaning "red feather", and on the right an example of a Demonstrative trial, meaning "that feather".

a debrief questionnaire at the end of the experiment, which verified that word meanings were learned accurately (we also return to this in the General Discussion).

2.1.2 Procedure

Participants were instructed that they would be learning part of a new language called Nápíjò, spoken by around 10,000 people in a rural region of Southeast Asia. All words and phrases were presented both auditorily and orthographically. The experimental session lasted about 15 minutes, and was divided into (1) Noun training, (2) Noun testing , (3) Noun-modifier training, (4) Noun-modifier testing, and finally (5) extrapolation to two modifiers.

Participants were first trained on the Nouns in the language. On each trial, participants saw an image of an object displayed on a table and were given its label in Nápíjò. They were instructed to click on the image to move on to the next trial. Noun training was composed of 15 trials (five trials for each of the three Nouns). This was immediately followed by 15 trials of Noun testing in which a picture appeared with two buttons beneath it, each containing a label. Participants were instructed to click the matching label. Feedback was given (button colour turned green or red, the correct description played regardless of response).

They were then trained on Noun-modifier combinations. Each trial had two parts. First, two images appeared, each illustrating one of the two modifiers for a given modifier type (e.g., "black" and "red", or "this" and "that"). A description of the first picture was provided, while the second picture was greyed out. Then, a description of the second picture was provided while the first was greyed out. Recall was tested immediately following this: The two pictures appeared again (in random order), and the description for one of them was given. Participants were instructed to click the picture matching the description. The first 12 such trials were blocked by modifier type, with random choice of which modifier type was introduced first (six trials per block), followed by a further 12 trials with both types intermixed (one trial for each Noun-modifier combina-

tion). Feedback was given after each trial (image background turned green or red, plus a beep sound if incorrect). Participants were then tested on their knowledge of the Noun-modifier combinations. Noun-modifier testing was composed of 24 trials. The foil labels for each picture were either an incorrect Noun *or* an incorrect modifier of the same type. On each trial, a picture appeared, with two potential descriptions below it. Participants were told to click on the matching description (12 Noun trials, 12 modifier trials, three for each modifier, random order). Feedback was given on each trial (button colour turned green or red, the correct description played, regardless of response).

In the critical testing phase, participants were tested (without training) on phrases with a Noun and *two modifiers*, using a forced-choice task, as in Culbertson & Adger (2014). On each trial a picture appeared, with two potential descriptions below it. Participants were told to click on the matching description (16 trials total, four for each modifier, random order). The two descriptions always included the correct lexical items, in postnominal order. They differed only in whether the order was homomorphic (e.g., N-Adj-Dem) or not (e.g., N-Dem-Adj). No feedback was given.

2.1.3 Participants

All participants were self-reported native English speakers recruited through Amazon's Mechanical Turk online recruiting platform and gave informed consent. They received 3.50 USD as compensation. We recruited a total of 99 participants who were randomly assigned to either the Dem-Adj, Dem-Num or Num-Adj condition. No participants were excluded for failing to reach at least 85% accuracy in the single modifier test trials. This was the same exclusion criterion used in previous studies. We thus analysed data from 99 participants: 33 in the Dem-Adj condition, 29 in the Dem-Num condition, and 37 in the Num-Adj condition.

2.2 Results

Results from Experiment 1 are presented in fig. 4. Following Culbertson & Adger (2014) and Martin, Ratitamkul, et al. (2019), we conducted our analyses using logistic mixed-effects models implemented in R (Bates 2014). Specifically, for each condition we designed full models with the binary dependent variable Homomorphic, along with by-participant random intercepts. We used likelihood ratio tests to compare these models to null models with no intercept term to see if on average participants chose homomorphic orders above chance level. We found a homomorphism preference in the Dem-Adj condition ($\beta = 1.15$, $SE = 0.37$, $\chi^2(1) = 8.77$, $p < 0.001$) but not in the Dem-Num ($\beta = 0.60$, $SE = 0.33$, $\chi^2(1) = 3.08$, $p = 0.08$) or Num-Adj ($\chi^2(1) < 1$) conditions.

2.3 Discussion

The results of Experiment 1 revealed a preference for homomorphic word orders, but only if the set of modifiers learned was Adjectives and Demonstratives. That is, participants preferred noun phrases with the order N-Adj-Dem (as in "mug red this") over the

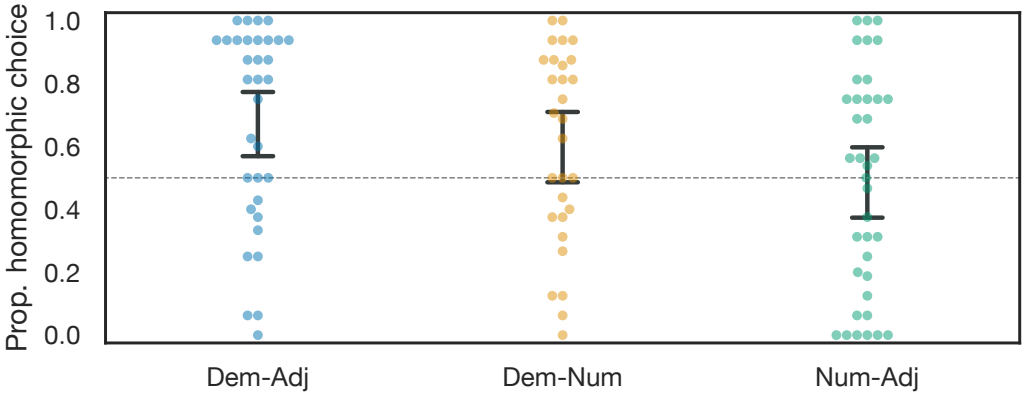


Figure 4: Proportion homomorphic preference in Experiment 1 (forced-choice task online) by condition. Each point represents an individual participant and error bars represent 95% confidence intervals calculated on participant means.

order N-Dem-Adj (as in "mug this red"). However, they showed no preference between orders N-Adj-Num (as in "mugs red two") and N-Num-Adj (as in "mugs two red"), nor any preference between orders N-Num-Dem (as in "mugs two these") and N-Dem-Num (as in "mugs these two"). We return to this asymmetry below. Recall that our paradigm was designed to be more naturalistic, with no orthographic representation of the native language, in order to reduce the likelihood of participants relying on an explicit "flipping" strategy. To verify this, we tabulated the number of participants reporting different response strategies, displayed in the left column of table 2. The most striking thing about the explicit reports participants provided is that not a single participant in our experiment reported "flipping" the phrases. Interestingly, there are also very few participants who report considering English (or any other language[6]) in determining their responses.[7]

To summarise, we successfully reduced participants' reliance on the problematic strategy that appeared to be in play in previous experiments. We found results that are partially convergent with previous work–an intriguing sensitivity to modifier type–but also a qualitatively weaker bias for homomorphic orders.[8] One possible explanation for the

[6] The most commonly reported L2s of participants across experiments were Spanish, French, and Chinese. None of these languages strongly predispose learners to postnominal homomorphism. While in both French and Spanish, Adjectives may occur postnominally, only Spanish allows postnominal modifiers of other kinds (namely, Demonstratives). Given that almost no participants referred to their L2 in their strategy reports, and that postnominal Demonstratives in Spanish are less frequent than their prenominal counterparts, we think that it is unlikely that L2 knowledge could explain our results.

[7] Two participants reported basing their responses on which modifier they learned first (included in Other in table 2). To verify that blocked learning did not explain a significant amount of the variance observed (perhaps without participants being aware of it), we designed a statistical model similar to those reported above, but with the predictor variable Block Order replacing Condition. Removing Block Order as a predictor did not significantly affect model fit ($\chi^2(1) < 1$), meaning we did not find evidence for block order affecting homomorphism preference, in line with the explicit strategy reports.

[8] We conducted an additional experiment similar to Experiment 1, but using four adjective words instead of just two. We obtained nearly identical results, providing further evidence for difference between modifier types. Those results are reported in Martin, Abels, et al. (2019).

Table 2: Counts of participants reporting different strategies in Experiment 1 ($N = 99$), Experiment 2 ($N = 86$) and Experiment 3 ($N = 61/91$) (cf. table 1). "Other" includes references to block order, data saving errors, or incomprehensible strategy reports. [*NB: the strategy reports for the Dem-Num condition of Exp. 3 are in a building currently locked down due to Covid-19*].

	Exp. 1	Exp. 2	Exp. 3
Flipping	0	0	0
(Non-)Homomorphic	29	29	43
L1 and/or L2	3	1	1
No strategy or no response	47	51	14
Other	20	4	3

overall weaker homomorphism preference here is that our results are a better representation of English speakers' underlying bias for homomorphism: it is present but not categorical for Adjectives and Demonstratives, and not present for Adjectives and Numerals or for Numerals and Demonstratives. The methodological confound—i.e., a salient flipping strategy—in previous studies may have simply masked this result by amplifying participants' choice of the homomorphic orders. This is consistent with another possibility, that the general task set up (shared with previous studies), in which ordering choices are presented orthographically, leads to a relatively weak preference. For example, by choosing, rather than generating the phrases themselves, participants may be less likely to fully acquire the lexical items (and thus their syntactic categories and meanings), and less likely to create a mental representation of the complex meanings they saw. Indeed, production tasks are known to encourage the construction of stronger mental representations (better retention) than perception tasks (Fawcett 2013). This may have led to a weaker influence of the underlying structure, or increased attention to linear order. Increased attention to linear order may even have encouraged some participants to follow the modifier order of English (although none explicitly stated this). Alternatively, it could be the case that having both options consistently available simply weakens what would otherwise be a stronger bias.

In Experiment 2, we attempt again to find evidence for a stronger, across the board homomorphism preference by replicating Experiment 1 using a production task. In this case, participants must *produce* phrases by speaking their responses into a microphone, rather than by choosing among a set of orthographically-presented options in the critical testing phase.

3 Experiment 2

In Experiment 2, we use the same artificial language stimuli as in Experiment 1, but here no orthographic representation of the language was given at any time during the experiment; stimuli were presented orally only. Participants thus had to produce all lexical items and phrases in the language orally.

3.1 Methods

3.1.1 Stimuli

The stimuli for Experiment 2 were similar to those for Experiment 1, with the exception of the tone and stress patterns used. Piloting revealed that the presence of tone was not simply ignored by the English speakers, as we had hoped, but rather served as a barrier to vocabulary learning. In order to make the stimuli sound less English-like, while still being produceable by English speakers, we opted to remove tone and rather place stress consistently on the final syllable of each stimulus word. The lexicon was thus as follows: two items that served as Adjectives or Numerals /puˈku/ and /taˈka/; two items that served as Numerals or Demonstratives /hiˈmi/, /hoˈno/; and three Nouns /eˈje/, /uˈhu/, and /iˈti/. Visual stimuli did not differ from Experiment 1.

3.1.2 Procedure

The procedure was based on Experiment 1, but was adapted for a production task.

Noun training was identical to Experiment 1, except participants advanced between trials by pressing the spacebar. As in Experiment 1, there were 15 Noun training trials (five for each noun). Noun training was followed by a 15-trial Noun testing phase that was identical to Experiment 1. Noun-modifier training was also identical to Experiment 1.

After participants were trained on the modifiers, they took part in their first recorded production phase. In single modifier production trials, an image from the Noun-modifier training phase was presented. Participants were then requested to produce the phrase themselves (recording was automatically controlled by the experimental programme), and to press the spacebar to save their recording and move on to the next trial. After each recording, the correct phrase was played back to participants as feedback. There were eight such trials.

Following this first production phase was another comprehension test. Participants saw two images and heard one Nápíjò phrase. They had to click on the image that corresponded to the phrase. There were a total of 24 such trials. Trials tested recall of either modifier or Noun labels. In trials that tested recall of Noun labels, one image was correct (e.g., a picture of a red ball) and the foil image displayed the correct modifier, but the incorrect Noun (e.g., a picture of a red feather). In trials that tested recall of modifier labels, one image was correct (e.g., a picture of a red ball) and the foil image displayed the correct Noun, but an incorrect modifier (e.g., a picture of a black ball) Noun and modifier trials were randomly intermixed.

This comprehension test was followed by another single modifier production phase, but with 16 total trials. This second single modifier production phase was used as an exclusion criterion during data analysis. Again, participants saw an image and had to produce the correct Nápíjò phrase. Feedback was identical to the previous production phase.

The final phase of the experiment was the critical two-modifier production phase which also consisted of 16 trials. Participants were presented with images that depicted

a Noun combined with two modifiers (e.g., two red balls) and were asked to produce the Nápíjò phrase they thought corresponded to the image. No feedback was given during this production phase.

3.1.3 Participants

All participants were self-reported native English speakers recruited through Amazon's Mechanical Turk online recruiting platform and gave informed consent. They received 4 USD as compensation. We recruited a total of 156 participants who were randomly assigned to either the Dem-Adj, Dem-Num or Num-Adj condition. Some of these participants did not successfully record audio responses and were thus not included in the analysis ($N = 23$). Participants were additionally excluded from analysis if they did not reach at least 85% accuracy in the single modifier production trials ($N = 36$),[9] or if they did not produce at least 10 analysable double modifier trials ($N = 25$). Note that some of these participants overlap, such that only 47 individual participants were excluded. We thus analysed data from 86 participants: 29 in the Dem-Adj condition, 29 in the Dem-Num condition, and 28 in the Num-Adj condition.

3.2 Results

Results from Experiment 2 are presented in fig. 5. The analysis of Experiment 2 was identical to that of Experiment 1. As in Experiment 1, we found a homomorphism preference in the Dem-Adj condition ($\beta = 10.93$, $SE = 2.20$, $\chi^2(1) = 43.86$, $p < 0.001$). In contrast to Experiment 1, we also found a homomorphism preference in the Dem-Num condition ($\beta = 9.01$, $SE = 1.80$, $\chi^2(1) = 33.36$, $p < 0.001$), as well as the Num-Adj condition ($\beta = 7.96$, $SE = 2.21$, $\chi^2(1) = 14.45$, $p < 0.01$). That is, we found above-chance homomorphism preferences in all conditions.

As in the previous experiment, we saw no evidence that participants relied on metalinguistic strategies like flipping or block order. Response strategies are reported in the middle column of table 2.

3.3 Discussion

In Experiment 2, participants were required to produce the language orally, rather than choose among orthographically-presented options. We predicted that this would strengthen the homomorphism preference across the board, and reveal a preference for N-Num-Dem and N-Adj-Num. This prediction was borne out: in all conditions there was a significant preference for homomorphic orders.

The critical difference between Experiments 1 and 2 is the manner in which participants gave their responses: In Experiment 1, participants chose from orthographically-presented response options, while in Experiment 2, participants had no linguistic prompt, and were required to produce themselves what they thought the best description of the

[9] That means that in the second single modifier production trials, they did not produce the correct items (either by entirely omitting a word or by confusing the lexical items) in at least three trials.

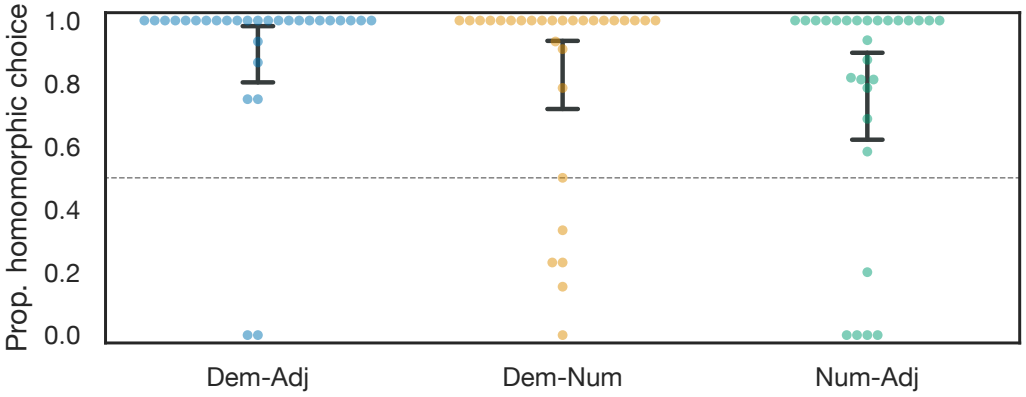


Figure 5: Proportion homomorphic preference in Experiment 2 (production task online) by condition. Each point represents an individual participant and error bars represent 95% confidence intervals..

image was in Nápíjò. As mentioned above, there are a number of mechanisms by which production could have resulted in a stronger homomorphism preference. First, participants had to acquire the lexical items to a higher standard in Experiment 2 since they had to themselves produce them. This could lead to a better representation of the meanings and/or syntactic categories of the items in Experiment 2. Second, producing phrases aloud might encourage formation or activation of the mental representation posited in (1) which is hypothesised to drive linear order preferences; activating this representation more strongly should lead to a stronger preference.[10] Third, participants were exposed to both homomorphic and non-homomorphic responses during the critical test phase of Experiment 1 (since both options were presented on screen). Thus, a preference for homomorphism might have been weakened by the mixed input that participants were receiving.

To confirm the across-the-board homomorphism preferences reported here, and to verify that the differences observed between Experiments 1 and 2 were driven by the manner in which participants gave their response, we conducted a third experiment. In Experiment 3, we used a procedure very similar to that of Experiment 2, but participants took part in a controlled laboratory version of the experiment.

[10] These first two arguments are bolstered by results from an additional experiment we ran online where participants were required to type, rather than produce their responses aloud. There, we found homomorphism preferences for Dem-Adj ($N = 35$, mean: 88.3%; $\beta = 11.05$, $SE = 2.01$, $\chi^2(1) = 53.24$ $p < 0.001$) and Dem-Num ($N = 32$, mean: 73.3%; $\beta = 8.92$, $SE = 1.76$, $\chi^2(1) = 19.39$ $p < 0.001$) combinations, though the numerical trend for Num-Adj combinations was not significantly above chance ($N = 36$, mean: 59.7%; $\beta = 1.69$, $SE = 1.38$, $\chi^2(1) = 1.68$ $p > 0.05$). For the sake of brevity, we do not report the full results of that experiment here.

4 Experiment 3

4.1 Methods

The stimuli and procedure for Experiment 3 were identical to those for Experiment 2 with a few exceptions. First, Experiment 3 was conducted in person in laboratory conditions, while Experiment 2 was run online. Second, participants were explicitly prompted to repeat the Nápíjò phrases even during training in order to prepare them for the recorded production phases. Finally, during the comprehension test between the first and second production phases, trials were randomly selected so that half of them tested recall of Noun labels and half tested recall of modifier labels.

4.1.1 Participants

Participants for Experiment 3 were self-reported native English speakers recruited through the University of Edinburgh's MyCareerHub and gave informed consent. They received 5.00 GBP as compensation. We recruited a total of 108 participants. Participants were randomly assigned to the Dem-Adj, Dem-Num, or Num-Adj condition. Participants were excluded from analysis if they did not reach at least 85% accuracy in the single modifier production trials ($N = 9$), or if they did not produce at least 10 analysable double modifier trials ($N = 13$). Note that some of these participants overlap, such that only 17 individual participants were excluded. We thus analysed data from 91 participants: 30 in the Dem-Adj condition, 30 in the Dem-Num condition, and 31 in the Num-Adj condition.

4.2 Results

Results from Experiment 3 are presented in fig. 6. The analysis of Experiment 3 was identical to that of Experiments 1 and 2. As in Experiments 1 and 2, we found a homomorphism preference in the Dem-Adj condition ($\beta = 9.23$, $SE = 1.79$, $\chi^2(1) = 28.26$, $p < 0.001$). In contrast to Experiment 1, but like Experiment 2, we also found a homomorphism preference in the Dem-Num condition ($\beta = 11.36$, $SE = 2.18$, $\chi^2(1) = 41.33$, $p < 0.001$), as well as the Num-Adj condition ($\beta = 2.56$, $SE = 0.91$, $\chi^2(1) = 9.16$, $p < 0.01$). That is, we again found above-chance homomorphism preferences in all conditions.

As in the previous experiments, we saw no evidence that participants relied on metalinguistic strategies like flipping or block order. Response strategies are reported in the right hand column of table 2.

4.3 Discussion

As in Experiment 2, we saw homomorphism preferences across the board in Experiment 3. This suggests that it is indeed the type of task (forced choice versus production) that explains the differences between our experiments. Below, we provide a series of post hoc analyses aimed at testing the influence of task type (forced choice vs. production testing)

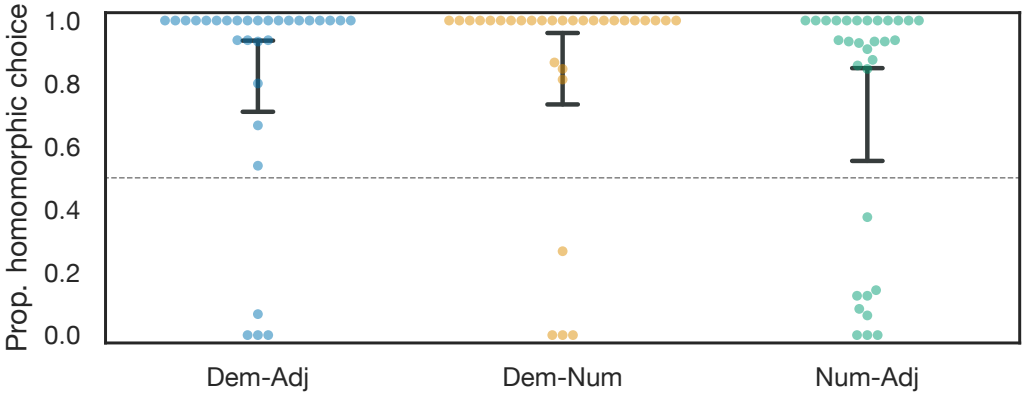


Figure 6: Proportion homomorphic preference in Experiment 3 (production task in the lab) by condition. Each point represents an individual participant and error bars represent 95% confidence intervals..

and population (online vs. lab) across all of our results. We also consider the asymmetry between modifier types discussed in Culbertson & Adger (2014).

5 Comparison across experiments

In this section, we provide a post hoc analysis of all of our data, in order to ascertain the robustness of ordering preferences across speaker populations and task types, and across modifier types.

Indeed, two key differences exist amongst our experiments: the population (MTurkers [Experiments 1 and 2] or in-person lab participants [Experiment 3]) and the task type (forced-choice [Experiment 1] or production [Experiments 2 and 3]). To test the relative influence of population and task type on our results, we designed a logistic mixed effects model with the binary dependent variable Homomorphic, a fixed effect of Population (MTurkers or lab participants, deviation coded), a fixed effect of Task (forced choice or production, deviation coded), and by-participant random intercepts, using pooled data from Experiments 1, 2, and 3. This model was compared to a model that excluded the factor Population and to a different model that excluded the factor Task using likelihood ratio tests. The full model was not found to significantly differ from the model excluding Population ($\chi^2(1) < 1$), but was found to be a significantly better fit of the data than the model excluding Task ($\beta = 2.47$, $SE = 0.84$, $\chi^2(1) = 9.05$, $p < 0.01$). This confirms that it is the task difference that explains the discrepancies in results across experiments, rather than the population tested.

Let us return now to the differences we saw amongst modifier types in Experiment 1: specifically, we saw a strong homomorphism preference for Demonstrative-Adjective combinations, but no preference for Demonstrative-Numeral or Numeral-Adjective combinations. In previous studies, similar numerical differences amongst modifier types were argued to reflect the fact that Dem and Adj are structurally more distant from one another than Dem and Num or Num and Adj (Culbertson & Adger 2014; Martin, Abels,

et al. 2019). In other words, assuming (1), there is more structure intervening between Dem and Adj than between Dem and Num, or Num and Adj. Under this account, non-homomorphic orders of Dem and Adj would then be less likely than non-homomorphic orders of Dem and Num or Num and Adj, because permuting structurally more distant modifiers is particularly dispreferred. Here, we provide a post hoc test of this hypothesis by comparing preferences for structurally more distant combinations (Dem-Adj) with structurally less distant combinations (Dem-Num and Num-Adj), again pooling the data from the three experiments. We designed a logistic mixed effects model with the binary dependent variable Homomorphic, a fixed effect of Distance (long or short, deviation coded), and by-participant random intercepts. We compared this model with a similar model that did not include the fixed effect using a likelihood ratio test. This model was found to significantly differ from the full model, indicating that we did observe an effect of structural distance on homomorphism preferences ($\beta = 1.33$, $SE = 0.58$, $\chi^2(1) = 5.38$, $p < 0.05$). That is, overall, homomorphism preferences for Dem-Adj combinations were significantly stronger than those for Dem-Num and Num-Adj combinations.

6 General discussion

We have reported results from three artificial language learning experiments testing the hypothesis that learners infer the order of nominal modifiers in a new language based on an underlying hierarchical structure rather than based on the surface statistics of their native language.

Based on this hypothesis, learners were predicted to use *homomorphic orders*, namely those that transparently reflect an independently motivated underlying structure in which Adjectives form a constituent with the Noun alone (i.e., [Adj [N]]), Numerals form a larger constituent including that one (i.e., [Num [Adj [N]]]), and finally, Demonstratives combine highest (i.e., [Dem [Num [Adj [N]]]]), as in (1) above. Homomorphic orders can be derived by linearising this structure, while non-homomorphic orders cannot.

While previous work has suggested that both English- and Thai-speaking participants show a homomorphism bias (Culbertson & Adger 2014; Martin, Ratitamkul, et al. 2019), methodological features of those experiments may have encouraged at least one metalinguistic strategy. This could have led participants to use a particular order without any structure-based bias in place. Specifically, these experiments presented native language phrases along with the artificial language options that participants were required to choose among. While native language words were originally used with the idea that this would provide the strongest evidence that learners do not rely on the surface (e.g., bigram) statistics of their language, we have shown above that this design led many participants to report "flipping" the native language words in order to get the artificial language order. Because both Thai and English orders are homomorphic, and Thai and English words were presented orthographically in these experiments, a metalinguistic strategy of reversing the order of the words necessarily pushes participants to choose homomorphic orders as well (e.g., Dem-Adj-N flips to N-Adj-Dem). The evidence from these previous experiments is therefore open to criticism.

Here, we addressed this criticism by designing full artificial languages that did not rely on direct translation between a speaker's L1 and the artificial language, replicating the results with an improved methodology. As in previous experiments, learners were taught a new language with postnominal modifiers, but were not given any evidence about the relative order modifiers when more than one was present. They were asked at test to either choose between a homomorphic or non-homomorphic order of phrases with multiple modifiers (Experiment 1), or to produce such phrases aloud (Experiments 2 and 3). In Experiment 1, we found a preference for homomorphic orders, but only for noun phrases containing a Demonstrative and an Adjective. In Experiments 2 and 3,where participants had to *produce* complex noun phrases aloud, we observed homomorphism preferences across the board, for Dem-Adj-N, Dem-Num-N *and* Num-Adj-N combinations.

Critically, in none of the experiments reported here did participants report using the type of metalinguistic flipping strategy which was commonly reported in previous experiments. In addition, during the debrief questionnaire, participants were asked to report translations of the words they had learned. Nearly without exception, participants correctly learned the Noun, Adjective, and Numeral meanings, reporting translations such as *cup*, *tea*, or *coffee* and *red*, *black*, *grey*, etc. The only variability came from the Demonstrative meanings. While most participants correctly reported translations like *proximal* or *distal*,[11] some participants reported meanings like *left* and *right*,[12] and a couple of participants reported meanings like *my* and *your*. To verify that theses alternative interpretations did not influence our results, we reran the analysis reported at the end of section 5 (i.e., considering structural distance), including only data from participants who gave proximal/distal translations in the Demonstrative conditions.[13] Again, we found stronger homomorphism preferences in the Dem-Adj condition (larger structural distance) than in the Dem-Num and Num-Adj conditions (smaller structural distance) ($\beta = 2.34$, $SE = 0.89$, $\chi^2(1) = 6.96$, $p < 0.01$). Our findings were therefore not driven by the subset of participants who failed to interpret the Demonstrative stimuli as intended. All in all, our results are the strongest experimental support to date of the role of hierarchical structure in driving noun phrase word order preferences.

Interestingly, the asymmetry amongst modifier types seen most clearly in Experiment 1, but confirmed in our post hoc analysis of our combined data mirrors numerical asymmetries reported in previous work—a seemingly stronger homomorphism preferences for noun phrases containing a Demonstrative and and Adjective. Culbertson & Adger (2014) appeal to the notion of structural distance to explain this asymmetry. Since Dem and Adj are more structurally distant than Dem and Num or Num and Adj, see (1), the bias

[11] These meanings were variously reported as *this* and *that* or *(over) here* and *(over) there*.

[12] The position of the girl in the Demonstrative illustrations was invariant: she was pointing to the left part of the table for the proximal meaning and to the right part of the table for the distal meaning. We chose this more consistent presentation so that the visual stimuli for Demonstrative trials would not be too complex or distinct from the other trials. To avoid left/right interpretations in future experiments, it would of course be possible to have the girl appear in various positions on Demonstrative trials.

[13] The average preference for homomorphism including all participants was 79.9% (±3.3%, standard error calculated on participant means) in the Dem-Adj condition and 76.5% (±3.5%) in the Dem-Num condition. When including only participants who correctly reported proximal/distal translations, these averages remained very similar: 82.3% (±5.3%) for the Dem-Adj condition and 77.5% (±6.1%) for the Dem-Num condition.

against producing a non-homomorphic order for the former might be stronger. While our experiments did not set out to test these asymmetries, it is worth considering whether they can shed further light on the interaction between structure and linear order.

The claim that structural distance can explain this asymmetry in learners' behaviour relies on the assumption that the full hierarchy in (1) is present or activated in some way in speaker-listeners' minds, even when one of the categories is not being expressed. For example, one could posit that in a noun phrase like "these green cars" (Dem-Adj-N), an empty Num (and its corresponding structure) is present. Indeed the theory of noun phrase word order proposed in Cinque (2005) does exactly that: every noun phrase is built from a large set of functional projections, each of which may or may not house a lexical item (e.g., Dem, Num, Adj, N) which is pronounced. However, it is notable that the vast majority of phrases containing (at least) one of these three modifier types have just a single modifier (83% in the English treebanks, compared to 14% with two modifiers, and 3% with three or more, Nivre et al. 2018). Further, almost all phrases with multiple modifiers (94%) have only a single modifier *type* (i.e., more than one Adjective). While this makes it all the more striking that speakers have such strong intuitions about the order of complex noun phrases (both in their native language and when learning a new one), it calls into question whether speakers would really represent all this structure each time they produce a noun phrase.

Another possibility is that our structural distance effect is actually driven by differences in dependency length which affect production or utterance planning. For example, if English speakers track dependencies between Nouns and nominal modifiers of different types, then they may have learned that the dependency between Noun and Dem can be longer while the dependency between Noun and Adj is always short. Permuting the order of Dem and Adj to produce a phrase like "feathers these red" would then result in a very local dependency being outside a normally more distant dependency. This effect would be present, but reduced, when producing a phrase like "feathers two red", where the local dependency is also lengthened, but more minimally. Again, the paucity of multiple modifiers mentioned above calls into question how such consistent preferences might be learned via dependency length differences.

Furthermore, this interpretation is not the typical way in which dependency length has been argued to affect word order. The typical claim is that languages tend to use orders that minimise dependency length (Hawkins 1990; Gibson 1998; Futrell, Mahowald & Gibson 2015). For example, in English "Mora threw out the trash" is generally preferred to "Mora threw the trash out" because the noun phrase dependent of the verb, "the trash", is longer than the particle dependent, "out". Placing the shorter dependent closer to the head has the result of minimizing the overall length (e.g., in words) between heads and dependants. This is argued to be advantageous for processing. However, in the noun phrases we are testing, all the dependants are of the same length, thus "feathers these red" and "feathers red these" don't differ in terms of overall dependency length. Nevertheless, recent work has argued that Nouns and Adjectives are typically more tightly connected than Nouns and Demonstratives (with Nouns and Numerals falling in between, Culbertson, Schouwstra & Kirby to appear). For example, particular pairs of Nouns and Adjectives are very common—e.g., "red wine" or "tall building"—while particular Nouns and Demonstratives have no special association with one another. Culbertson,

Schouwstra & Kirby (to appear) quantify this in terms of pointwise mutual information (an information-theoretic measure of strength of association), and show that across many languages, Adjectives and Nouns are more closely associated on average than Numerals and Nouns, which are in turn more closely associated than Demonstratives and Nouns. It could be that the strength of this kind of dependency determines how likely speakers are to separate a head and dependent (see also Futrell, Levy & Gibson 2017; Hahn et al. 2018). Alternatively these differences in strength of association could simply be another reflection of the semantic or conceptual structure that underlies the hierarchical syntax representation in (1). For example, Culbertson, Schouwstra & Kirby (to appear) argue that the strength of association they measure likely reflects how objects relate to their properties, numerosities, etc *in the world*, not just in linguistic utterances.

We leave to future work the task of determining which of these explanations underlies the asymmetry between modifier combinations we find. However, note that this question is closely tied to whether the constituency structure posited in (1) is universal or learned (our evidence is compatible with both positions). The answer to this question could come from detailed study of a language with N-Dem-Num-Adj order, like Kîîtharaka, (Kanampiu 2017; Muriungi 2008), where, given the surface order, Dem could in principle compose with the Noun before Adj. If this is the case, then we would expect syntactic differences between English and Kîîtharaka–for example in the ellipsis constructions used to motivate the constituency of English. We would also predict speakers of Kîîtharaka to show a preference for orders which are *not* homomorphic to the structure in (1). By contrast, if the noun phrase word order in Kîîtharaka is derived from the same universal underlying structure as English, then we predict that even these speakers, who have a lifetime of experience with a non-homomorphic surface order, may still infer homomorphic order in a new language. We plan to test this in future work.

7 Conclusion

Abstract hierarchical representations, encoding relations among linguistic sub-units, are a key feature of many theories of syntax and semantics. In the noun phrase, such representations have been argued to have a universal form, in which Adjectives are structurally closest to the Noun, then Numerals, then Demonstratives (Cinque 2005; Abels & Neeleman 2012). This hierarchy can explain (among other things) a striking asymmetry in the frequency of noun phrase word order patterns among the world's languages: in the vast majority, Adjectives are *linearly* closer to the Noun, and Demonstratives farthest away (e.g., N-Adj-Num-Dem or Dem-Num-Adj-N). In other words, there appears to be a transparent mapping between the hypothesised structure and linear order in most languages. Building on previous research using artificial language learning experiments, here we aimed to provide robust experimental evidence that human learners are biased in favour of such transparent mappings. We showed that English speakers taught a new language in which modifiers were postnominal (unlike English), used their knowledge of the underlying structure to infer linear order. For example, they assumed that the language would have N-Adj-Dem order rather than N-Dem-Adj order. The latter is more similar to the surface order of English, but the former transparently reflects an underlying structure

in which Adj is structurally closer to the Noun than Dem. Importantly, these findings provide stronger evidence than those reported in previous research (e.g., Culbertson & Adger 2014), in which a potential confound—use of a metalinguistic flipping strategy—could have explained learners' preferences. We also found that the strength of the preferences we found in our experiments correlated with structural distance; learners' preferences for the relative order of Adj and Dem were very strong, while for Adj and Num and Num and Dem they were relatively weaker. These results show that, when learning a new language, learners' are sensitive to the detailed hierarchical structure of the noun phrase rather than the linear order of their native language. Our results also point to the need for future work exploring whether this structure is learned from experience with a particular language, or reflects a universal underlying representation that shapes noun phrase word order typology.

Ethics and consent

Ethics approval for all experiments was obtained from the PPLS Ethics Committee prior to data collection and informed consent was obtained from all participants.

Funding information

This work was supported by the Economic and Social Research Council [Grant number ES/N018389/1].

Acknowledgements

We would like to thank Rattanasuwan Rawan for translating the Thai strategy reports and Danielle Naegeli for running Experiment 3 and coding the strategy reports.

Competing interests

The authors have no competing interests to declare.

References

Abels, Klaus. 2015. Syntax. In Natalie Braber, Louise Cummings & Liz Morrish (eds.), *Exploring language and linguistics*, chap. 6, 137–167. Cambridge University Press.

Abels, Klaus & Ad Neeleman. 2012. Linear asymmetries and the LCA. *Syntax* 15(1). 25–74.

Adger, David. 2003. *Core syntax*. Oxford: Oxford University Press.

Bates, Douglas M. 2014. *Lme4: Mixed-Effects Modeling with R*.

Chomsky, Noam. 1965. *Aspects of a theory of syntax*. Cambridge: MIT Press.

Cinque, Guglielmo. 2005. Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry* 36(3). 315–332. JSTOR: 4179327.

Culbertson, Jennifer & David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences* 111(16). 5842–5847. https://doi.org/10.1073/pnas.1320525111.

Culbertson, Jennifer, Marieke Schouwstra & Simon Kirby. to appear. From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language*.

Dryer, Matthew S. 2018. On the Order of Demonstrative, Numeral, Adjective and Noun. *Language*.

Elbourne, Paul. 2008. Demonstratives as individual concepts. *Linguistics and Philosophy* 31(4). 409–466.

Ernst, Thomas. 2002. *The syntax of adjuncts*. Cambridge: Cambridge University Press.

Fawcett, J. M. 2013. The production effect benefits performance in between-subject designs: a meta-analysis. *Acta Psychologica* 142(1). 1–5.

Frank, Stefan L. & Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22(6). 829–834.

Futrell, Richard, Roger Levy & Edward Gibson. 2017. Generalizing dependency distance: comment on "Dependency distance: A new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews* 21. 197–199.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33). 10336–10341.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68(1). 1–76.

Gleitman, Lila R & Elissa L. Newport. 1995. Language. In Lila R Gleitman, Daniel R Osheron & Mark Liberman (eds.), *An Invitation to Cognitive Science*. Cambridge: MIT Press.

Goldberg, Adele. 2006. *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Greenberg, Joseph H. 1963. *Universals of language*. Cambridge: MIT Press.

Hahn, Michael, Judith Degen, Noah D Goodman, Dan Jurafsky & Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. In Tim T Rogers, Marina Rau, Jerry Zhu & Chuck Kalish (eds.), *Proceedings of the 40th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21(2). 223–261.

Heim, Irene & Angelika Kratzer. 1998. *Semantics in generative grammar*. Oxford: Blackwell Publishing.

Hurford, James R. 1987. *Language and number: the emergence of a cognitive system*. Oxford: Basil Blackwell.

Kanampiu, Patrick Njue. 2017. *The syntax of the determiner phrase in Kĩîtharaka, a Bantu language spoken in Kenya*. Chuka University Master's thesis.

Krifka, Manfred. 1995. Common nouns: A contrastive analysis of English and Chinese. In G N Carlson & Pelletier F J (eds.), *The generic book*, 398–411. Chicago: Chicago University Press.

Martin, Alexander, Klaus Abels, David Adger & Jennifer Culbertson. 2019. Do learners' word order preferences reflect hierarchical language structure? In Ashok K. Goel, Colleen M. Seifart & Christian Freksa (eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2303–2309. Montreal, Canada: Cognitive Science Society.

Martin, Alexander, Theeraporn Ratitamkul, Klaus Abels, David Adger & Jennifer Culbertson. 2019. Cross-lingusitic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*.

McCauley, Stewart M & Morten H Christiansen. 2017. Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science* 9(3). 637–652.

Muriungi, Peter Kinyua. 2008. *Phrasal movement inside Bantu verbs: Deriving affix scope and order in Kîîtharaka*. University of Tromsø Doctoral dissertation.

Nivre, Joakim et al. 2018. *Universal Dependencies 2.2*.

Partee, Barbara H. 1987. Noun phrase interpretation and type-shifting principles. In Jeroen Groenendijk, Dick de Jongh & Martin Stokhof (eds.), *Studies in discourse representation theory and the theory of generalized quantifiers*, 115–143. Dordrecht: Foris.

Piantadosi, Steven T. & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756. https://doi.org/10.1111/cogs.12088.

Rijkhoff, Jan. 1990. Explaining word order in the Noun Phrase. *Linguistics* 28(1). 5–42.

Rijkhoff, Jan. 2004. *The noun phrase*. Oxford: Oxford University Press.

Steddy, S. & V. Samek-Lodovici. 2011. On the ungrammaticality of remnant movement in the derivation of Greenberg's Universal 20. *Linguistic Inquiry* 42(3). 445–469.

Steedman, Mark. 2018. A formal universal of natural language grammar. Manuscript, University of Edinburgh.

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