Do learners' word order preferences reflect hierarchical language structure?

Do learners' word order preferences reflect hierarchical language structure?

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

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

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

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

Abstract

Previous research has argued that learners infer word order patterns when learning a new language based on knowledge about underlying structure, rather than linear order (Culbertson & Adger, 2014). Specifically, learners prefer typologically common noun phrase word order patterns that transparently reflect how elements like nouns, adjectives, numerals, and demonstratives combine hierarchically. We test whether this result still holds after removing a potentially confounding strategy present in the original study design. We find that when learners are taught a naturalistic "foreign" language, a clear preference for noun phrase word order is replicated but for a subset of modifier types originally tested. Specifically, participants preferred noun phrases with the order N-Adj-Dem (as in "mug red this") over the 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"). We interpret this sensitivity as potentially reflecting an asymmetry among modifier types in the underlying hierarchical structure.

Keywords: language; learning; syntax; typology

Introduction

A large body of work has claimed that sensitivity to abstract hierarchical structure drives the acquisition of syntax (e.g., Chomsky, 1965). At the same time, there is evidence to suggest that language learners track surface-level statistics, including co-occurrence patterns among words (e.g., Saffran, Aslin, & Newport, 1996). In a recent paper, Culbertson and Adger (2014) used a pseudo-artificial language learning task to argue that learners privilege abstract structural relations among words to linear order when they learn syntactic features of a new language. Moreover, they suggest that sensitivity to these structural relations—which in their case pertain to noun phrase word order—can explain a well-studied typological generalisation, known to linguists as Universal 20 (Greenberg, 1963). In the current paper, we highlight some potential methodological issues with the paradigm used by Culbertson and Adger (2014), and test whether their finding is replicated once the paradigm is improved.

Research in generative syntax posits an underlying hierarchical structure for the noun phrase: [Dem [Num [Adj [N]]][1] (Adger, 2003; Cinque, 2005; Abels & Neeleman, 2012). In this hierarchy, which can be interpreted as reflecting semantic or conceptual structure, the adjective forms a constituent with the noun to the exclusion of the numeral and demonstrative; that sub-constituent combines with a numeral, and the resulting unit combines with a demonstrative to make a larger constituent. The structure provides a straightforward explanation for why, in most languages, adjectives are placed linearly closest to the noun, while demonstratives are furthest away (e.g., Dryer, 2018). For example, in English *these two red cars*, in Thai (the equivalent of) *cars red two these*. Both these orders can be read directly off the underlying structure, while others, like N-Dem-Num-Adj cannot. While such orders can in principle be derived by movement, they are rarely found. Culbertson and Adger (2014) refer to orders like Dem-Num-Adj-N and N-Adj-Num-Dem (as well as any other order that can be read directly off of the structure [Dem [Num [Adj [N]]]) as *isomorphic*—they preserve an isomorphic relation between the proposed underlying hierarchical structure and the surface linearisation.

Culbertson and Adger (2014) sought to provide evidence that learners are sensitive to this underlying structure, and use it to infer word order, rather than simply copying the linear order in their native language. To show this, they taught English speakers simple noun phrases in a pseudo-artificial language, with English words, but non-native-like word order. Participants saw an English phrase like *red shoe*, and were taught it would be *shoe red* in the new "language"; similarly *this car* would be *car this*. Participants were subsequently shown phrases with multiple modifiers, like *this red car*, and asked to guess the relative order of post-nominal modifiers in the language. The authors reason that if learners' inferences are guided by their knowledge of surface-level features of English, they should guess the non-isomorphic order (i.e., *car*

[1] Abbreviations: N(oun) (e.g., *car*), Adj(ective) (e.g., *red*), (Num)eral (e.g., *two*), Dem(onstrative) (e.g., *this*).

this red), which has its modifiers in English order. By contrast, if their inferences are guided instead by knowledge of the abstract structure described above, then they should infer the isomorphic order (i.e., *car red this*). Participants in their experiment overwhelmingly inferred isomorphic orders, suggesting sensitivity to the hypothesised universal structure rather than surface statistics of English.

While this result is intriguing, the paradigm used by Culbertson and Adger (2014) is unusual in several respects. First, even relative to other work using artificial language learning paradigms, this task is very non-naturalistic. Second, the task may encourage a particular strategy. Specifically, English phrases along with their "translations" in the language–also English words–were presented visually. Participants may have adopted an explicit strategy of reversing or "flipping" the English words to determine their responses. For example, during training participants could relate a translation like *shoe red* to the English phrase *red shoe* shown onscreen by reversing the words. Using the same strategy to guess the correct two-modifier phrase translation would then mean flipping the English *this red shoe* to *shoe red this*. Here, we aim to determine whether the apparent bias for isomorphic orders reported by Culbertson and Adger (2014) is replicated using a standard artificial language learning task, with a more naturalistic, completely novel language.

Experiment 1

The experiments we report on in the present paper are part of a larger cross-linguistic comparison project. We followed the methodology reported by White et al. (2018) and designed artificial languages using only sounds contained in all of the languages we plan to test. The phonological inventory of our artificial languages was thus reduced to five vowels, and a small set of voiceless (non-aspirated) stops, nasals, and the voiceless glottal fricative, all shared by the languages we plan to test in.[2] The languages all have lexical tone (for planned experiments with speakers of tonal languages), though the tones do not serve to contrast words from one another (thus the English-speaking participants can simply ignore them). As in Culbertson and Adger (2014), we taught participants phrases with a noun and a single modifier (either and adjective and a demonstrative, or an adjective and a numeral), and then asked them to guess the relative order of modifiers when both were present. Crucially, in contrast with Culbertson and Adger (2014), we used completely novel stimuli and did not present written L1 equivalents of the phrases participants were learning. This was done to reduce the possibility, present in Culbertson and Adger (2014), that participants would simply "flip" L1 word orders to translate into the artificial language they were learning.

Methods

Stimuli The artificial language had five lexical items. There was a single noun meaning *feather*, represented by the label /jè/. There were two adjectives (meaning *red* and *black*), and two items that served as either demonstratives (*this* and *that*) or numerals (*two* and *three*) depending on the condition the participant was assigned to. Labels for these modifier classes were created in pairs:/púkù/, /tàká/ and /hìmí/, /hónò/. The two pairs of stimuli were randomly assigned to be either adjectives or demonstratives/numerals. We privileged within-pair similarity (so /púkù/ and /tàká/ both contain only voiceless stops for example) to facilitate the learning process.[3] 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.

Visual stimuli were pictures of simple cartoon scenes. Objects (always feathers) 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, feathers were drawn in light grey; feathers 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. 1.

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-modifier training, (3) noun-modifier testing, and finally, (4) extrapolation to two modifiers. Participants were first trained on the (single) noun in the language. On each trial, participants saw the object and were given its label in Nápíjò. They were instructed to click on the image to move on to the next trial. There were five such trials. 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 was given. Participants were instructed to click the picture matching the description. The first eight such trials were blocked by modifier type, with random choice of which modifier type was introduced first (two trials per modifier), followed by a further 16 trials with

[2]Experiment 3 contains some additional fricatives that will not be used with non-English-speaking populations.

[3]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.

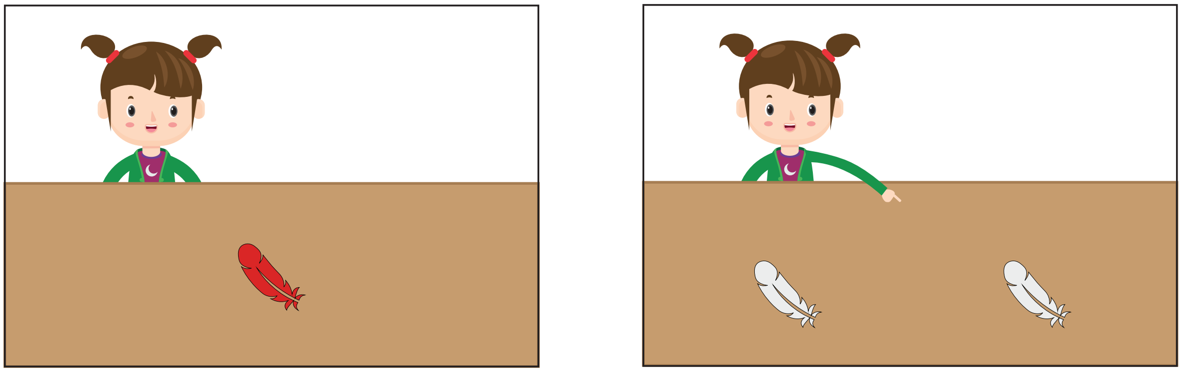

Figure 1: Single modifier trial visual stimuli examples. On the left, an example of an adjective trial, meaning "red feather", and on the right an example of a demonstrative trial, meaning "that feather".

both types intermixed. 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. 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 foil description always included a modifier of the same type. 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*. 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 post-nominal order. They differed only in whether the order was isomorphic (e.g., N-Adj-Dem) or not (e.g., N-Dem-Adj). No feedback was given.

Participants All participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 70 participants who were randomly assigned to either the Demonstratives or Numerals condition. A total of eight participants were excluded (four in each condition) because they failed to reach at least 85% accuracy in the single modifier test trials (this is the same exclusion criterion reported by Culbertson and Adger (2014)). We thus analysed data from 35 participants in the Demonstratives condition and 27 in the Numerals condition.

Results

Following the analyses reported in Culbertson and Adger (2014), we analysed, for each condition, whether participants demonstrated an average preference for isomorphic orders on two modifier trials. Results from Experiment 1 are presented on the lefthand side of fig. 2. All analyses were performed by implementing logistical mixed-effects models in the `lme4` package in R (Bates, 2014). We designed full models with the binary dependent variable Isomorphic along with by-participant random effects. We used likelihood ratio tests to compare these models to null models with no intercept term to see if on average participants chose isomorphic orders above chance level. We found no isomorphic preference in either the Demonstratives ($\chi^2(1) < 1$) or the Numerals conditions ($\chi^2(1) < 1$).

Discussion

Contrary to Culbertson and Adger (2014), we did not observe any preference for isomorphic order in our artificial language learning task. However, given that our methodology differed in a number or respects from the original studies (and replications), we considered possible explanations for our null result. First, Culbertson and Adger (2014) used English words in their experiment, whereas we used nonce words. It is therefore worth verifying that participants in our experiment interpreted the words as intended. In a debrief questionnaire, participants were asked to report the meanings of the words they had learned. Participants invariably reported correct translations for adjectives (colour words) and numerals. However, meanings given for demonstratives and nouns varied to some degree. For demonstratives, most participants reported translations such as *this* and *that*, or *here* and *there*. Both these translations are consistent with a demonstrative interpretation: although *here* and *there* are sometimes called adverbs, their meaning and syntax are similar to *this* and that, and indeed they are the demonstrative words in many languages (Diessel, 2006). However, some participants gave responses such as *left* and *right* (indeed, the absolute and relative positions were confounded in our stimuli). The variation in interpretation of the demonstrative may have weakened the results to some degree. However, the interpretation of the noun suggests a more obvious issue.

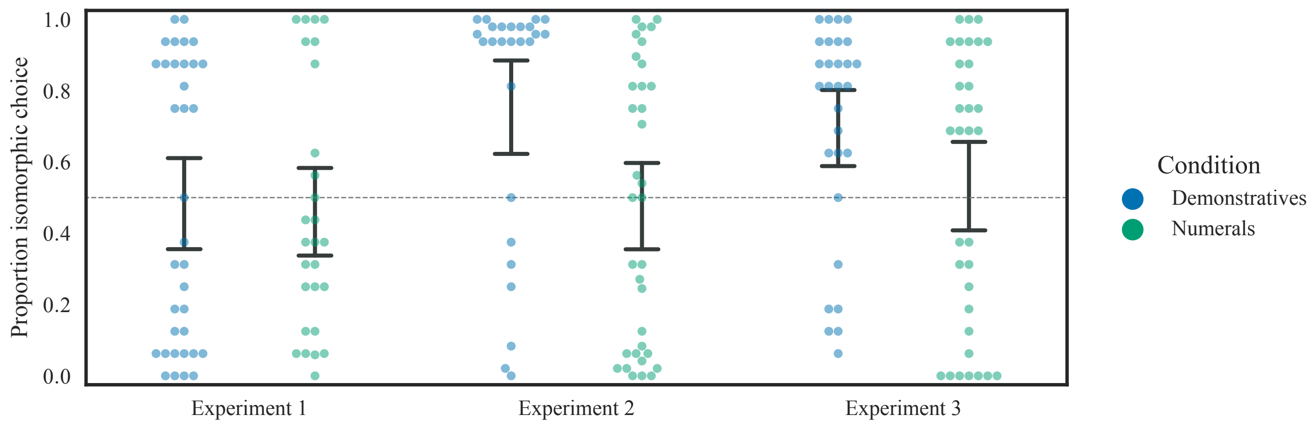

Figure 2: Proportion isomorphic preference in each experiment by condition. Each point represents an individual participant and error bars represent 95% confidence intervals.

While some participants correctly reported the English word *feather* for the word /jè/. Some, did not report a translation at all, suggesting they may not have understood its meaning. Others reported incorrect meanings, giving function words like *the* as translations. Note that the set size of the nouns differs dramatically from Culbertson and Adger (2014), where 20 nouns were used: we used a single noun in Experiment 1. Given that every trial always consisted of /jè/ + *x*, it was therefore possible for participants to completely discount that word (or indeed treat it as a determiner). This suggests the possibility that participants may not have been treating our stimuli as noun phrases (i.e., not attending to the noun head), but simply as strings of modifiers. If so, participants may have adopted any number of response strategies. For example, they could have interpreted the strings as copulative sentences (e.g., "this (one) (is) red"), or simply strings of modifiers. In either case, they would not have learned the intended Noun-modifier structure of the language. In Experiment 2, we therefore expanded the set of nouns in the artificial lexicon. We hypothesised that variability in the noun would cause participants to treat our stimuli as noun phrases, resulting in a preference for isomorphic orders.

Experiment 2

Methods

Stimuli The stimuli for Experiment 2 were similar to those in Experiment 1; only the nouns differed. We created audio and visual stimuli for three objects (*feather*, *ball*, *mug*) which were assigned the names /éjè/, /úhù/, and /ítì/, respectively. All modifier stimuli were identical to Experiment 1.

Procedure The procedure was identical to Experiment 1 except the initial training and testing phases were slightly lengthened. 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 labels beneath it. Participants were instructed to click the matching label. Feedback was given (button colour turned green or red, the correct description played regardless of response). Noun-modifier exposure was composed of 12 trials blocked by modifier type (six trials per block, two for each noun-modifier combination), followed by an additional intermixed block of 12 trials (one trial for each noun-modifier combination). Noun-modifier testing was composed of 24 trials (two trials for each noun-modifier combination). The foil labels for each picture were either an incorrect noun *or* an incorrect modifier of the same type. Finally, for the critical test phase, a random set of 16 trials was constructed for each participant.

Participants As in Experiment 1, all participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 71 participants who had not participated in Experiment 1. Participants were randomly assigned to either the Demonstratives or Numerals condition. A total of 11 participants were excluded (seven in the Demonstratives condition and four in the Numerals condition) because they failed to reach at least 85% accuracy in the single modifier test trials. We thus analysed data from 26 participants in the Demonstratives condition and 34 in the Numerals condition.

Results

Results from Experiment 2 are presented in the middle of fig. 2. The analysis of Experiment 2 was identical to that of Experiment 1. We found an isomorphic preference in the Demonstratives condition ($\beta = 2.25$, $SE = 0.60$, $\chi^2(1) = 11.35$, $p < 0.001$) but not in the Numerals condition ($\chi^2(1) < 1$).

Discussion

The results of Experiment 2 revealed a preference for isomorphic 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 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"). Interestingly, this asymmetry has been reported numerically in all previous experiments on isomorphism. As discussed above, Culbertson and Adger (2014) found statistically significant isomorphism preferences for all pairs of modifiers (adjective, numeral and demonstrative), and when all three modifiers were present (not tested here). However, they report a numerical difference among the groups such that the isomorphism preference is strongest with adjective and demonstrative. Indeed, they cite this as further evidence that English speakers are sensitive to the underlying hierarchical structure, since adjectives and demonstratives are structurally more distant than adjectives and numerals (or numerals and demonstratives). In a lab replication of the original study (which was conducted on Mechanical Turk), A. Martin, Ratitamkul, Abels, Adger, and Culbertson (in press) replicated both the general isomorphism preference and the difference among modifier pairs. They also report a replication with Thai speakers, whose L1 order is N-Adj-Num-Dem. These speakers were trained on an artificial language with prenominal modifiers, and they then inferred prenominal isomorphic orders like Dem-Adj-N in the critical two-modifier test phrase. There again the same difference among modifier pairs was present. These studies report only numerical differences. Our findings therefore present the clearest evidence yet that the isomorphism preference may be sensitive to modifier type.

Nevertheless, we did not replicate an isomorphism preference for the Numerals condition. Additionally, the isomorphism preference found for the Demonstratives condition is (numerically) weaker than reported in these previous studies. By design, we have reduced the likelihood that participants are relying on an explicit "flipping" strategy, and we have made the language itself more naturalistic. Thus, one possibility is that our results are a better representation of English speakers' underlying bias for isomorphism: it is present, but not categorical for adjectives and demonstratives, and not present for adjectives and numerals. We return to this in the general discussion. There is, however, one other major difference between our experiment and previous experiments which could plausibly weaken or mask an isomorphism preference, namely the relative size of the modifier categories. In both Culbertson and Adger (2014) and A. Martin et al. (in press), the relative class sizes approximately match what one would typically find in a natural language: largest set size for adjectives, then numerals, and a small set of demonstratives.[4] In our experiments, all modifier classes contained two elements. In Experiment 3, we test the possibility that using a more naturalistic relative size for the modifier classes might amplify the isomorphism preference, perhaps revealing the isomorphism preference between numerals and adjectives reported in previous work.

[4]For example, 694 adjective vs. 172 numeral, 5 demonstrative types among all noun phrases in the English Universal Dependencies Treebank (Nivre et al., 2017).

Experiment 3

Methods

Stimuli The stimuli for Experiment 3 were similar to those for Experiments 1 and 2. The only difference was in the number of adjectives. Specifically, four adjectives were created (/tàkás/, /pùkúf/, /kàpáθ/, and /kùtúʃ/) and mapped to four colour meanings ("black", "red", "blue", and "green", respectively). Visual stimuli similar to those in Experiments 1 and 2 were also created.

Procedure The procedure was identical to Experiment 2 except for the following: Noun-modifier training was all blocked (in order to balance frequency of exposure to each combination without increasing the number of trials too much). Each block was composed of 12 trials. In the adjective block, each adjective was shown once with each noun. In the numeral or demonstrative block, each modifier was shown twice with each noun. The noun-modifier testing block was slightly longer than in Experiment 2, with 36 trials total (2 trials for each noun-modifier combination). No changes were made to the critical two modifier testing phase (again, 16 trials total, randomly constructed). Note that the frequency of exposure to each modifier class was the same, only the number of elements in each class differed.

Participants As in Experiments 1 and 2, all participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 76 participants who had not participated in Experiment 1 or Experiment 2. Participants were randomly assigned to either the Demonstratives or Numerals condition. A total of 13 participants were excluded (nine in the Demonstratives condition and four in the Numerals condition) because they failed to reach at least 85% accuracy in the single modifier test trials. We thus analysed data from 29 participants in the Demonstratives condition and 34 in the Numerals condition.

Results

Results from Experiment 3 are presented on the right-hand side of fig. 2. The analysis of Experiment 3 was identical to that of Experiments 1 and 2. As in Experiment 2, we found an isomorphic preference in the Demonstratives condition ($\beta = 1.24$, $SE = 0.36$, $\chi^2(1) = 10.37$, $p < 0.01$) but not in the Numerals condition ($\chi^2(1) < 1$).

Discussion

In Experiment 3, we tested whether the isomorphism preference found in Experiment 2 would be amplified, and extended to the Numerals condition if the relative sizes of the modifier classes were more naturalistic. This was not borne out; rather we replicated the findings of Experiment 2: an isomorphism preference for noun phrases with a demonstrative and an adjective, but not for noun phrases with a numeral and an adjective. This finding therefore reinforces the asymmetry reported in Experiment 2, and the numerical patterns reported in both Culbertson and Adger (2014) and A. Martin et al. (in press). In the next section, we investigate statistically the general pattern of results across experiments described here.

Comparison across experiments

Two manipulations distinguished Experiments 1, 2, and 3. First, the size of the noun class. In Experiment 1, participants learned only one noun, while in Experiments 2 and 3 they learned three. Second, the size of the adjective class. In Experiments 1 and 2, participants learned only two adjectives, while in Experiment 3 they learned four. We thus performed an analysis considering these two binary variables, included in our models using contrast coding. This allowed us to explore the interaction between these two factors and the factor Condition in one single statistical model. The model predicted Isomorphic order choice from three fixed binary factors: Condition (Demonstratives or Numerals), Noun Class Size (one noun or three), and Adjective Class Size (two adjectives or four). We also included interactions between Condition and Noun Class Size and between Condition and Adjective Class Size as well as by-participant random intercepts. We then designed reduced models each excluding one factor or interaction, and compared them to the full model (again using likelihood ratio tests).

We found that removing Noun Class Size significantly worsened the model fit ($\beta = 1.08$, $SE = 0.49$, $\chi^2(1) = 4.70$, $p < 0.05$). This indicates that participants who learned an artificial language with three nouns showed a stronger isomorphism preference than those who learned an artificial language with only one noun. We also found that removing the interaction between Condition and Noun Class Size significantly worsened the model fit ($\beta = -2.65$, $SE = 0.98$, $\chi^2(1) = 7.10$, $p < 0.01$). This confirms our observation that amongst the participants who learned artificial languages with three nouns, those in the Demonstratives conditions showed an isomorphism preference while those in the Numerals conditions did not. Removing the factors Adjective Class Size ($\chi^2 < 1$) and Condition ($\chi^2 = 1.46$, $p = 0.23$) did not worsen the model fit, nor did removing the interaction between Condition and Adjective Class Size ($\chi^2 = 1.40$, $p = 0.24$).

General discussion

This paper aimed to test the preferences of English speakers learning about the noun phrase word order of a new language. Previous research using a pseudo-artificial language learning paradigm reported a strong preference for so-called isomorphic noun phrase orders, like N-Adj-Dem or N-Adj-Num, which transparently reflect the hypothesised hierarchical structure of the noun phrase: [Dem [Num [Adj [N]]]] (Culbertson & Adger, 2014; A. Martin et al., in press). This has been claimed to show that speakers' inferences about a new language are not based on the surface linear order of their native language, but on a (potentially universal) underlying hierarchical structure. Moreover, the results suggest the possibility that a preference for orders which are isomorphic to this structure might explain why these orders overwhelmingly outnumber non-isomorphic orders in the typology (Cinque, 2005; Abels & Neeleman, 2012; Dryer, 2018).

We sought to replicate these findings using an improved methodology, designed to address the possibility that the original results reflected the availability of an explicit strategy which may have encouraged participants to choose isomorphic orders by visually flipping the English words. We used a standard artificial language learning paradigm, with a relatively more naturalistic language. In Experiment 1, we used a minimal vocabulary, with only a single noun, and found no isomorphism preference. In Experiment 2, we added additional nouns to encourage participants to treat stimuli as noun phrases. Here, we found an isomorphism preference for phrases including a demonstrative and an adjective, but not for phrases including a numeral and an adjective, an asymmetry which mirrors numerical differences reported in earlier studies. In Experiment 3, we attempted to strengthen the isomorphism preference by making the number of words in each modifier category more naturalistic (in terms of relative size). This did not change the results, but rather again revealed that learners' isomorphism preference was sensitive to the modifier categories involved.

Importantly, our results show that in a more naturalistic artificial language learning task, where participants are unlikely to use an explicit strategy of flipping English words to determine order in the new language, an isomorphism preference is still found. Some confirmation that participants are not using a simple flipping strategy in our experiments comes from self-reports given at the end of the task. Of the 185 participants that were retained for data analysis in our three experiments, only one referred to a flipping strategy in the debriefing questionnaire. Instead, common strategies included "no strategy", "I just went with my gut feeling" (67 such reports), or simple descriptions of their order choices like "I placed colour words closer to the object name, then numbers" (50 such reports). This contrasts starkly with the strategies reported by participants in Culbertson and Adger (2014)'s study. We recovered the data from that study and analysed the 89 participant strategy reports from their Experiment 1: 47 of them reported some kind of explicit flipping-based strategy (compared to only 11 "no strategy"). Our replication of their effect with a more naturalistic artificial language is thus an important contribution to this line of research.

Our results also highlight the persistent difference between modifier types, found numerically in earlier experiments, and confirmed statistically here. While it is possible that something about our task is still masking a (weaker but present) isomorphism preference for numerals and adjectives, there is some reason to suspect that the *asymmetry* at least is real. In fact, using the data collated by Dryer (2018), we can observe that non-isomorphism between numerals and adjectives, or numerals and demonstratives is more common crosslinguistically (35 and 64 languages respectively) than non-isomorphism between adjectives and demonstratives (27 languages). This may reflect the fact that adjectives and demonstratives are more distant from one another in terms of underlying hierarchical structure.

As mentioned in the introduction, this hierarchy can be conceived of as reflecting semantic composition, or conceptual structure. Indeed linear order patterns more generally have been argued to reflect both (Rijkhoff, 1990; Baker, 1985; Bybee, 1985; Rice, 2000). One possibility is that the underlying hierarchy of nominal modifiers reflects differences in conceptual closeness (or inherentness) between particular modifier types and nouns (Kirby, Culbertson, & Schouwstra, 2018; Culbertson, Schouwstra, & Kirby, under revision). Under this account, adjectives are conceptually closest to nouns because they are more likely to reflect inherent properties of individual nouns (e.g., colour, size, texture, etc). Numerals are typically less closely linked with particular nouns (though some clearly are, e.g., four seasons, seven days of the week). Demonstratives, being deictic elements, are by their nature not associated with particular nouns. If the underlying hierarchical structure reflects these different conceptual relations between elements, then a preference for isomorphism is a preference to hierarchically cluster elements that are more closely related conceptually. Perturbing this preference would then be less costly when it involves elements that differ less in their conceptual closeness to the noun (e.g., Adj and Num), compared to elements that differ quite a lot (e.g., Adj and Dem) (for similar arguments about the relative order of adjectives, see J. E. Martin, 1969; Bouchard, 2002).

To summarise, the experiments reported here aimed to replicate the preference for isomorphic ordering in the noun phrase, first reported in Culbertson and Adger (2014). Using a more naturalistic artificial language learning task, we find that English speakers infer isomorphic orders of demonstrative and adjective. However, we found no evidence of an isomorphism preference for numerals and adjectives. Above we suggest one possible explanation for the difference between these two conditions: assuming that English speakers can either use an isomorphic order, or an order that reflects the surface linear order of their language, they are more likely to go with the latter when this would involve two modifiers that are more similar to each other, either in terms of structural distance, or in terms of conceptual closeness with the noun. That said, learners' sensitivity to the distribution features of the language (e.g., in Experiment 1) leave open the possibility that future experiments will reveal this bias as weaker but still present.

Acknowledgements

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

References

Abels, K., & Neeleman, A. (2012). Linear asymmetries and the LCA. *Syntax*, *15*(1), 25-74.

Adger, D. (2003). *Core syntax*. Oxford: OUP.

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

Bates, D. M. (2014). *Lme4: Mixed-effects modeling with R.*

Bouchard, D. (2002). *Adjectives, number and interfaces: Why languages vary*. Amsterdam: Elsevier.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry*, *36*(3), 315-332.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *PNAS*, *111*(16), 5842-5847.

Culbertson, J., Schouwstra, M., & Kirby, S. (under revision). *From the world to word order: the link between conceptual structure and language.*

Diessel, H. (2006). Demonstratives, joint attention, and the emergence of grammar. *Cognitive linguistics*, *17*(4), 463–489.

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

Greenberg, J. H. (1963). *Universals of Language*. MIT Press.

Kirby, S., Culbertson, J., & Schouwstra, M. (2018). The origins of word order universals: evidence from corpus statistics and silent gesture. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii).* NCU Press.

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

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, *8*(6), 697–704.

Nivre, J., Agić, , Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., ... Zhu, H. (2017). *Universal Dependencies 2.1.* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rice, K. (2000). *Morpheme order and semantic scope: Word formation in the Athapaskan verb*. Cambridge: Cambridge University Press.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., ... van de Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning. In *Proceedings of NELS 48.*