Replication of Culbertson and Adger (2014): Language learners privilege structured meaning over surface frequency

Sebastian Schuster
Department of Linguistics, Stanford University
sebschu@stanford.edu

Introduction

Linguistic universals, i.e., properties that hold for all human languages of the world, can be a great source of insights on the human language faculty. Properties that are true of every language potentially demonstrate limitations of which kind of languages can be learned and therefore give us insight into the cognitive processes of language acquisition. For this reason, linguistic universals play an important role in the Chomskyan program of uncovering a universal grammar. On of the primary goals of this research is to uncover properties and limitations of the human language faculty by investigating which common properties exist across languages and which kind of properties are not exhibited by other languages. One problem with this approach is, however, that it becomes more and more evident that there exist no non-trivial properties that hold for all languages (*absolute* universals, following Greenberg's terminology) but rather that there are just certain properties that are exhibited by a large number of unrelated language (*statistical* universals) (Evans & Levinson, 2009). While statistical universals cannot be used to back up any claims about limitations of human language learning processes, they are still interesting to cognitive scientists as they might be a result of certain (potentially innate) cognitive biases.

One well-known universal is Greenberg's Universal 20 (Greenberg, 1963) on the order of nominal modifiers:

> In pre-nominal position the order of demonstrative, numeral, and adjective (or any subset thereof) conforms to the order DEM-NUM-ADJ.
>
> In post-nominal position the order of the same elements (or any subset thereof) conforms to either the order DEM-NUM-ADJ or to the order ADJ-NUM-DEM.

While this universal is formulated as an absolute universal, later work has shown that at least 14 of the 24 possible orderings of these three modifiers and the noun appear in the languages of the world (Hawkins, 1983; Cinque, 2005). Nevertheless, there is a tendency for languages to exhibit one of the two orderings DEM-NUM-ADJ-N and N-ADJ-NUM-DEM (Cinque, 2005). This raises the question whether the predominance of these two orders is just accidental or whether there exists a cognitive bias to favor these two orderings.

Culbertson and Adger (2014, henceforth C&A) note that the scope of these three modifiers is universally assumed to be DEM $>$ NUM $>$ ADJ, which can be hierarchically structured as in the tree in Figure 1. Incidentally, if one linearizes this tree by assuming that modifiers are either all left-branching or all right-branching, one ends up with exactly the two predominant word-orderings. This observation leads C&A to their main research question: Do language learners make use of this structural information when making inferences about word orders in a new language or do language learners predominantly rely on surface frequency information?

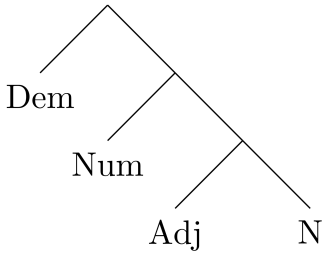

Figure 1: Hierarchical structure of the scope of noun phrases

C&A explore this question in two artificial language learning experiments. Their artificial language is exactly like English with the exception that nominal modifiers appear after the noun. The idea behind their experiment is to teach participants that nominal modifiers always appear after the noun in this language and solely based on this knowledge ask participants to "translate" phrases with multiple modifiers and consequently to make an inference about their order. C&A argue that participants could use either of the following two strategies to infer the order in the target language:

H1: Participants rely on surface frequency information from English which, for example, tells them that the bigram *these two* is much more likely than the bigram *two these* and therefore, they use the same order of modifiers as in English.

H2: Participants make use of structural information and therefore translate noun phrases such that the translation has a transparent scope-isomorphic order.

Based on the assumption that participants will learn that modifiers should appear in post-nominal position, these two hypotheses make the following predictions P1 and P2.

P1: When asked to translate an English noun phrase with multiple modifiers, participants select a translation in which the modifiers appear post-nominally in the same order as in English.

P2: When asked to translate an English noun phrase with multiple modifiers, participants select a translation in which the modifiers appear post-nominally in scope-isomorphic order, i.e., in the reverse order of English.

The implicit linking hypothesis here is that selecting a translation of an English phrase into this artificial language gives us insights about how syntactic information is represented in adults and how these representation affect the learning of new languages. C&A are a bit vague about whether this should tell us anything about first language acquisition or only about second language acquisition – they motivate this study primarily on the basis of research in first language acquisition but then (rightfully) only discuss the implications of their results on our understanding of adult representations and second language learning.

Experiment

My experiment is a replication of experiment 1 by C&A. In this experiment, participants are first exposed to a set of training trials in which they see an English phrase of the form MODIFIER NOUN and listen to its translation of the form NOUN MODIFIER in which the lexical items were the same but the word order differed. Each participant is exposed to only two types of modifiers, with the types depending on the condition (NUM and ADJ, ADJ and DEM, and NUM and DEM, for conditions 1, 2, and 3, respectively). The purpose of the training trials is to teach participants that modifiers should appear after the noun. The training trials are followed by a set of test trials in which participants are asked to translate English phrases of the form MODIFIER NOUN and MODIFIER1 MODIFIER2 NOUN. The two-word trials serve as control trials to test whether participants learned that modifiers should appear in post-nominal position. The three-word trials are the critical trials in which participants have to make an inference on the ordering of modifiers.

Methods

Participants I recruited 96 participants (32 per condition) with an US IP address through Amazon Mechanical Turk. Participants took on average between 7 and 8 minutes to complete the experiment and I paid them $1.65 for their participation.

Materials I used the same list of 30 nouns, 10 adjectives, 10 numerals, and 4 demonstratives as C&A which are listed in Table 1. C&A do not mention whether they allowed all combinations of adjectives and nouns, but as some of them such as *furry banana* are semantically anomalous, I limited the combinations to meaningful adjective-noun pairs, which excluded 83 of the 300 possible combinations.[1] I included all possible 300 numeral-noun pairs and 120 demonstrative-noun pairs. For the critical test trials, I included the same adjective-noun combinations with any numeral or determiner, and I used all numeral-demonstrative-noun triplets which agreed in number (excluding ungrammatical phrases such as *these one cherries*). I generated the spoken translations with the US English model of the Google Text-to-Speech API using the gTTS[2] package.

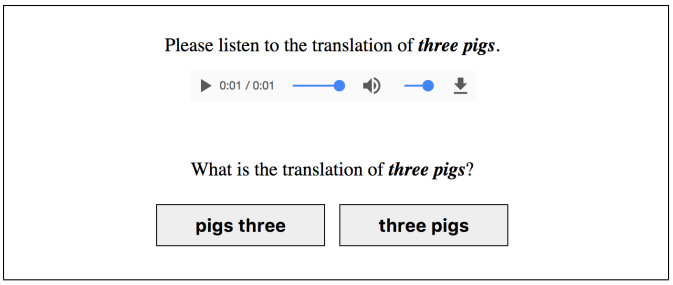

Figure 2: Example training trial from condition 1.

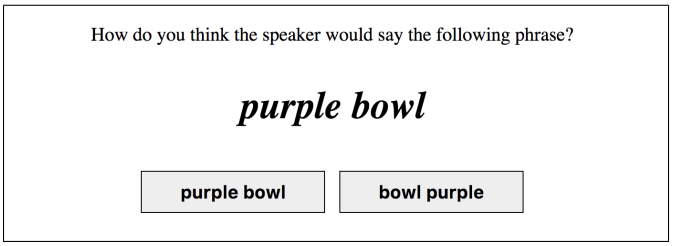

Figure 3: Example control trial from condition 1.

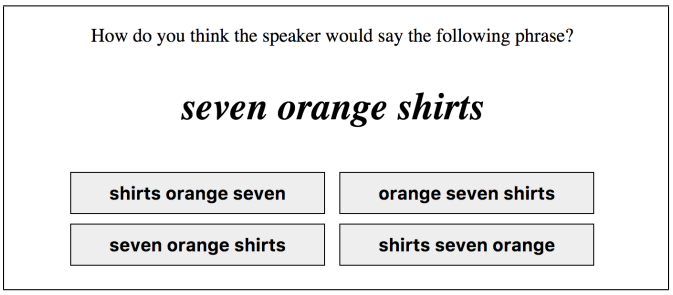

Figure 4: Example critical trial from condition 1.

Procedure For the training phase, participants were instructed that they would see a short phrase and listen to a translation into a language which uses the same words as English but "differs a bit from English". Training consisted of 30 trials, presented in random order: 15 trials with MODIFIER 1 NOUN phrases and 15 trials with MODIFIER 2 NOUN phrases. For each trial, participants saw a two-word phrase and heard a spoken translation of this phrase. After they listed to the translation, participants were asked to click on a button corresponding to what they heard. The translation automatically started playing after a 500ms delay and the buttons to select a translation only appeared after participants had listened to the translation. Participants could replay the translation as often as they wanted. An example training trial is shown in Figure 2.

[1] See https://github.com/sebschu/ling-245/blob/master/experiments/stimuli/stimuli.txt for the complete list of used modifier-noun pairs.

[2] https://github.com/pndurette/gTTS

Nouns	Adjectives	Numerals	Demonstratives
Banana, bowl, box, car, chair, cherry, couch, cow, dart, duck, handbag, hat, horse, key, lamp, leaf, pear, pig, pillow, pineapple, pitcher, ribbon, scarf, shirt, shoe, shovel, spatula, suitcase, tray, vase	Blue, green, orange, purple, dirty, furry, spotted, wooden, stone, striped	One, two, three, four, five, six, seven, eight, nine, ten	This, that, these, those

Table 1: List of nouns, adjectives, numerals, and demonstratives used to construct the stimuli.

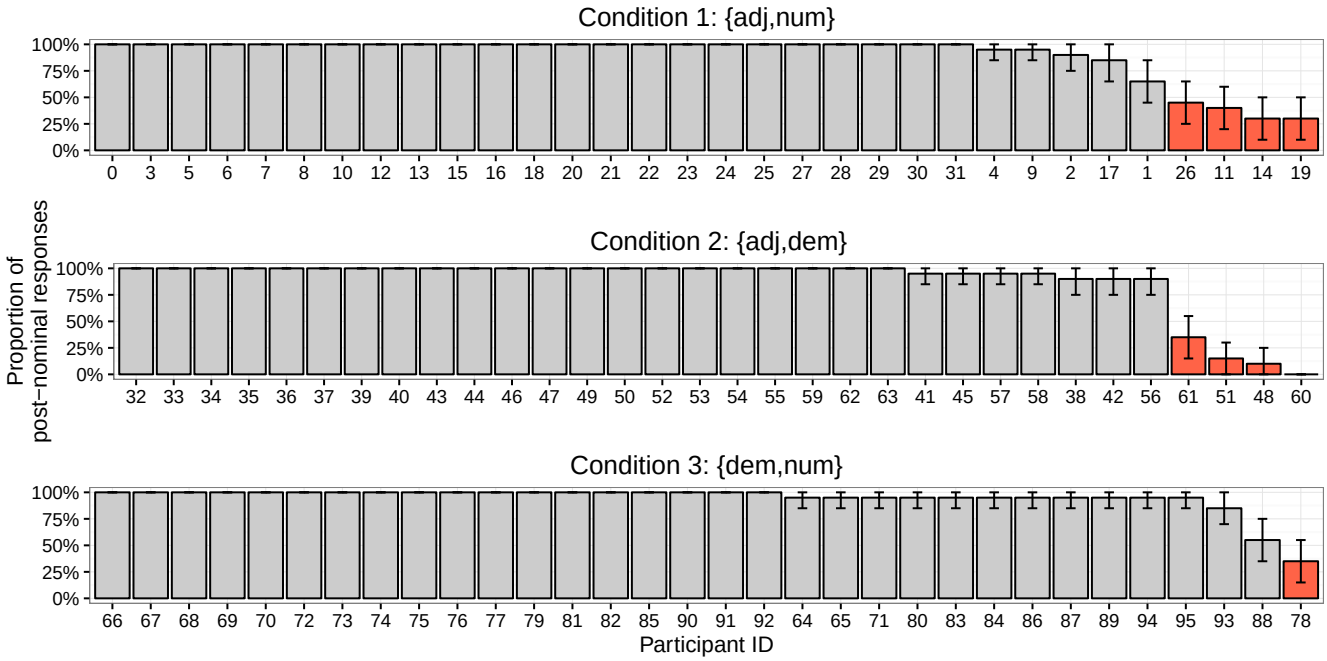

Figure 5: Proportion of post-nominal responses among control trials for each participant. Participants who selected the post-nominal phrase in less than 50% of the control trials and who I subsequently excluded from further analyses are highlighted in red.

After completing the training trials, participants saw another set of instructions asking them to translate the phrases that were presented in the following trials by clicking on the appropriate button. They were instructed that they might have heard some of the phrases during training but that there will also be new ones for which they should guess what they think the speaker would have said. The test phase consisted of 50 trials, presented in random order: 10 control trials with MODIFIER1 NOUN phrases, 10 control trials with MODIFIER2 NOUN phrases, and 30 critical test trials with MODIFIER1 MODIFIER2 NOUN phrases. Example control and critical trials are shown in Figures 3 and 4.

After completing all trials, participants were asked about their native language and any other languages that they had learned at some point in their life.

Results

Data coding and exclusions I coded all responses in which participants were selected the post-nominal scope-isomorphic ordering as 1, and all other responses as 0. Figure 5 shows the the proportion of post-nominal responses among control trials for each participant. Following C&A, I discard the data from all participants who consistently failed to select post-nominal responses. C&A do not mention a threshold for exclusion, so I excluded all 9 participants who selected the post-nominal translation in less than 50% of the control trials from further analyses.

Analysis C&A used a Bayesian mixed-effects logistic regression model for their analysis but as the description of their model is very terse, I was not able to reconstruct their exact model. For this reason and because I did not see an advantage of using a Bayesian model for this kind of analysis, I used a regular mixed-effects logistic regression model. My dependent variable is whether a participant selected the post-nominal scope-isomorphic translation and my only fixed effect is the condition, which I code using dummy coding with condition 1 being the reference condition.[3] Unlike C&A, I only include random intercepts for subjects but not for items because on the one hand, the model with random subject and item intercepts does not fit the data significantly better according to a likelihood ratio test ($\chi^2(1) = 0$, $p = 0.99$), and on the other hand, there are in total 1,737 different items among my 2,610 observations, so there is just one data point for many of the items and therefore it is unlikely that a model with random item intercepts is able to produce reliable estimates for these parameters.

The proportions of post-nominal scope-isomorphic responses for each condition across all participants are shown in Figure 6. This figure suggests that in all three conditions, participants selected the post-nominal scope-isomorphic response more often that would be expected by chance. This

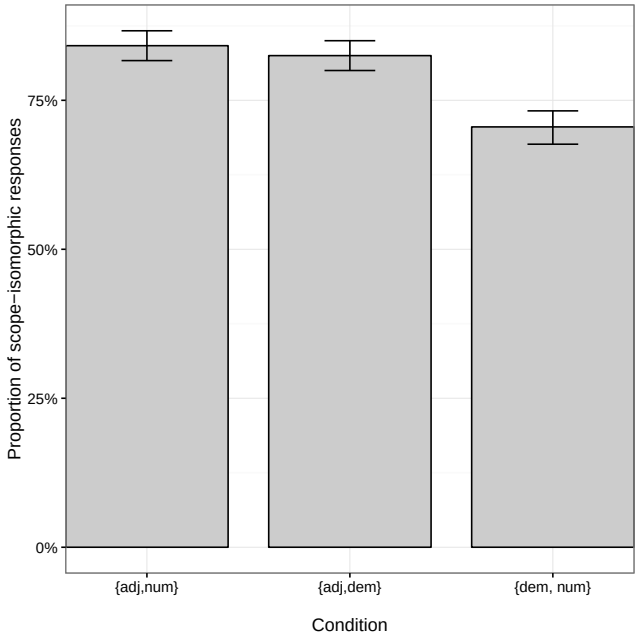

Figure 6: Proportions of post-nominal scope-isomorphic responses for each condition.

observation is confirmed by the mixed-effects model: the log odds of the scope-isomorphic response in the reference condition 1 are significantly above 0 ($\beta_0 = 3.3$, $z = 4.74$, $p < 0.001$), and the log odds in condition 2 and condition 3 do not significantly differ from the log odds in condition 1 (condition 2: $\beta_1 = 0.42$, $z = 0.43$, $p = 0.67$; condition 3: $\beta_1 = -1.61$, $z = -1.75$, $p = 0.081$). In summary, in all three conditions, participants were significantly more likely to select the post-nominal scope-isomorphic response than any other response and there are no significant differences between conditions.

Discussion

Similarly as C&A, I observed that participants predominantly selected the scope-isomorphic ordering in the critical trials and my effect size was similar to the one of C&A. This is potentially evidence for H2 and might indicate that participants indeed make use of innate or learned structural information when they decide on a word order in another language. However, as also pointed out by C&A, one problematic confound is that the scope-isomorphic order is also excactly the reverse English order and it could well be that participants just assume that the other language exhibits a reverse order compared to their native language. C&A dismiss this explanation on the basis of statistically significant differences between conditions 1 and 2 and 1 and 3. They argue that if participants just reversed the order, then it would be unlikely that there were any differences across conditions as this strategy should be completely independent of the types of modifiers. They further argue that pattern of a higher proportion of

[3] C&A used effect coding but as the differences between condition 1 and condition 2 and between condition 1 and condition 3 are the main differences we are interested in (as I explain in the discussion below), dummy coding seems to be a better choice here.

scope-isomorphic responses in condition 1 than in the other two conditions makes sense as the distance in terms of scope is larger in condition 1 than in the other two conditions and therefore it is less likely that people select translations whose order is not scope isomorphic.

All of these arguments are plausible but the big issue is that this effect does not seem to be reproducible. C&A run another experiment in which they teach all participants all three types of modifiers and have them make inferences on all three pairs of modifiers, so effectively each participant is exposed to all three conditions. In this within-subject experiment, they do not observe any differences across conditions. This potentially indicates that the differences across conditions in C&A's experiment 1 (a between-subject experiment) are just a random effect of some participants not paying attention. This is also in line with my results; I also did not observe statistically significant differences across conditions.

However, while the difference between conditions 1 and 3 in my experiment is not significant at a level of $\alpha = 0.05$, it cannot be dismissed that there is at least qualitatively a difference between these two conditions and that this difference is trending towards significance. A qualitative analysis of the individual participant's performance showed that this difference was mainly caused by a few participants who didn't show a clear strategy in translating the critical test phrases – they sometimes picked the scope-isomorphic order and at other times they picked the English order (see Figure 7 in the appendix). In the other two conditions, however, participants were more consistent. There might be two random factors for these differences. First, two of the inconsistent participants in condition 3 took much longer or much shorter than the average participants which might be a indication that they did not pay attention. Second, I ran condition 3 on a national holiday (Memorial Day) which might have also had a negative effect on some participants' attention.

But even if this difference is meaningful, my results still show a different pattern than the results of experiment 1 in C&A. For this reason, I am not convinced that the likely non-linguistic strategy of just reverting the word order can be dismissed based on the results. Consequently, it seems as if we cannot draw any conclusions about linguistic representations and their effects on second language learning.

One potential way to get rid of this problematic confound might be to consider a language with more flexible adjective order. Several Romance languages such as Spanish and French have both pre- and post-nominal adjectives. One could adapt the experiment that I ran for this replication to these languages and limit the adjectives for the training trials to those who typically appear pre-nominally and then ask participants to translate phrases with post-nominal adjectives. If participants consistently selected translations with the reverse order than the original phrase, i.e., placing the post-nominal adjective in front of the noun and all other modifiers behind the noun, then this would be evidence for the strategy of reversing the word order of one's native language. On the other hand, if participants consistently produced the scope-isomorphic ordering, i.e., leaving the adjective in place and moving the numeral or demonstrative to the end of the phrase, this would be evidence for a structural bias.

In conclusion, I was able to partially replicate the results by C&A but based on the absence of differences across conditions, the results appear to be equally compatible with Hypothesis 2 as well as the hypothesis that participants are making use of a non-linguistic strategy. For this reason, it seems unclear whether the present data allows us to draw any conclusions on adult syntactic representations and second language learning and therefore it seems problematic to derive at the conclusion of C&A that language learners make use of structural information when learning a second language.

References

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. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429.

Greenberg, J. H. (1963). *Universals of Language* (J. H. Greenberg, Ed.). Cambridge, MA: MIT press.

Hawkins, J. (1983). *Word Oder Universals*. New York, NY: Academic Press.

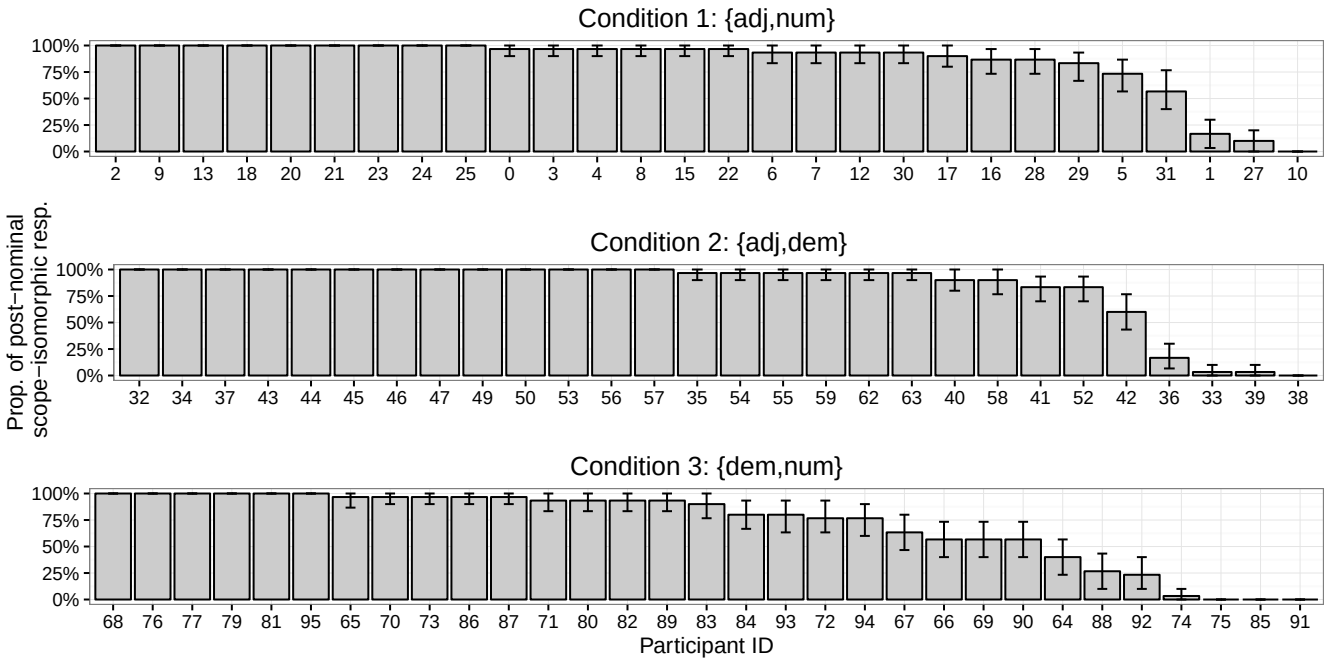

Figure 7: Proportion of scope-isomorphic post-nominal responses among critical test trials for each participant.