Appendix 2: pilot experiments

1. PIlot Experiment 1: learnability

1.1. Methodology

1.1.1. Participants

Participants were 80 Amazon Mechanical Turk workers who were compensated financially for their time ($6 per task). The task was advertised through mTurk as an “Alien Language Game.” All were native English speakers according to self-report, though some spoke other languages. Users were assigned randomly to one of four syntactic order conditions, as discussed below.

1.1.2. Stimuli

Stimuli were as in Experiment 1, except that the arrays were generated randomly for each participant rather than using a fixed set for all. Only four syntactic conditions were used: center-embedding and branching with either head-initial or head-final syntax (same as experiment 1, minus crossed syntax).

1.1.3. Procedure

Procedure for training and comprehension trials was the same as in Experiment 1 but without adposition training, and without training to criterion on single objects.
The production task was as in experiment 1, except that free text entry was used rather than buttons for the words (figure 1)

Figure 1: Write-in trial at level 3, pilot experiment 1

After completing the task, participants were asked to write in how they interpreted each adposition.

1.1.4. Analysis

Analysis was as in Experiment 1, except for the following: As there were only two structure conditions, sum-coded contrasts rather than Helmert contrasts were used in the models for analysis of comprehension and production correctness.

Write-in production task data was encoded and analyzed as follows: First, each word in every participant’s trial input was “spell-checked” against their given lexicon of nouns and adpositions, correcting each word to the one closest by normalized optimal string alignment (OSA) distance (R stringdist), wherein each one-letter addition, deletion, substitution or transposition of adjacent characters counted as one change. This measure was normalized by dividing the OSA distance by the length of the longer compared string. If no word in the lexicon had an OSA smaller than .75 from the input word, or if the word was identified as one of several English words inputted in place of adpositions by some users, this word was encoded as X.

Then the corrected participant production was processed through the same analyses as in Experiment 1, to determine accuracy and best match grammar.

1.2. Results

1.2.1. Comprehension

Complete statistical analysis of all measures can be found in Appendix 2, section 1. In the comprehension task (Figure 2), performance improved slightly with level in all conditions except center-embedded/head final, but this was not significant ($p > .09$). Improvement was greater in the head-initial condition than in the head-final condition ($B= -0.392$, $SE = 0.185$, $p=0.034$). Center-embedding, however, did not produce a significant effect on performance change with level ($p > .15$).

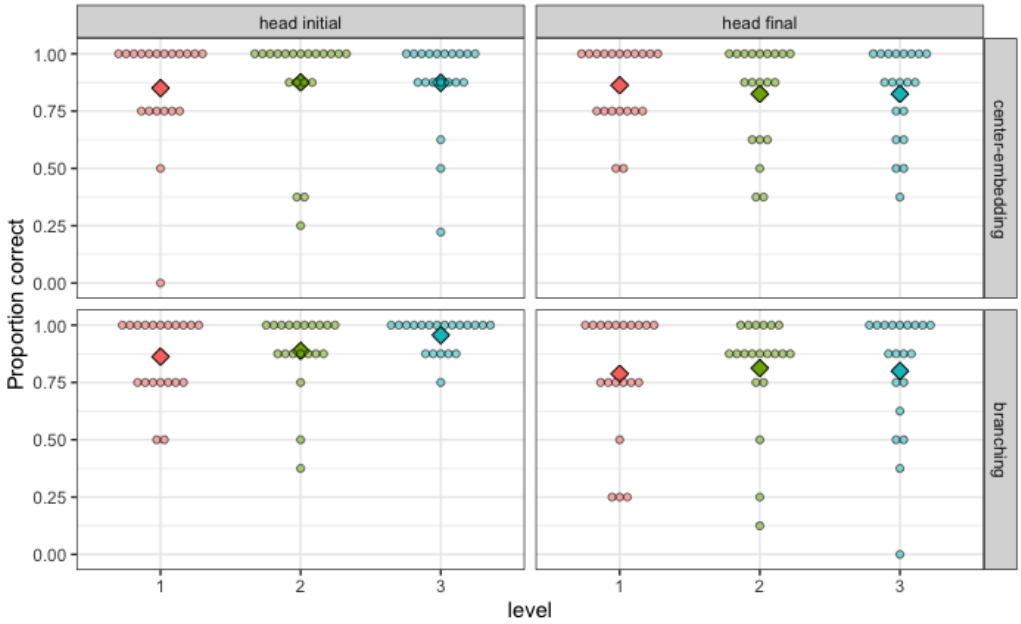

Figure 2: Performance on comprehension task, by condition and level. Large diamonds represent mean performance across all participants.

1.2.2. Production

In the production task (Figure 3), performance declined significantly with each level ($B = -.05$, $SE = .006$, $p < .001$) over all conditions, and there was a significant main effect of head order ($B= -0.072S$, $SE= 0.023$, $p =0.002$) but not of syntax type ($B= 0.008$, $SE = 0.023$, $p = 0.718$). The decline in performance with level was significantly less in the center-embedding conditions ($B= 0.02$, $SE = 0.006$, $p=0.003$), and greater in the head-final conditions ($B = -0.022$, $SE= 0.006$, $p= 0.001$). In addition, there was a significant interaction between head order and syntax type ($B = -.05$, $SE = .023$, $p = .025$), i.e. performance was worse overall in the center-embedded/head-final condition. For levels where both seen and unseen arrays appeared in the writing task (2 and 3 items), production accuracy was compared across previously seen and unseen. Performance was slightly, not significantly worse on unseen arrays ($B = -.019$, $SE = .014$, $p = .166$), while there was a significant interaction between unseen and head-final order ($B = -.036$, $SE = .014$, $p = .009$). Interaction with center-embedding was not significant ($B= -0.004$, $SE = 0.014$, $p = 0.781$).

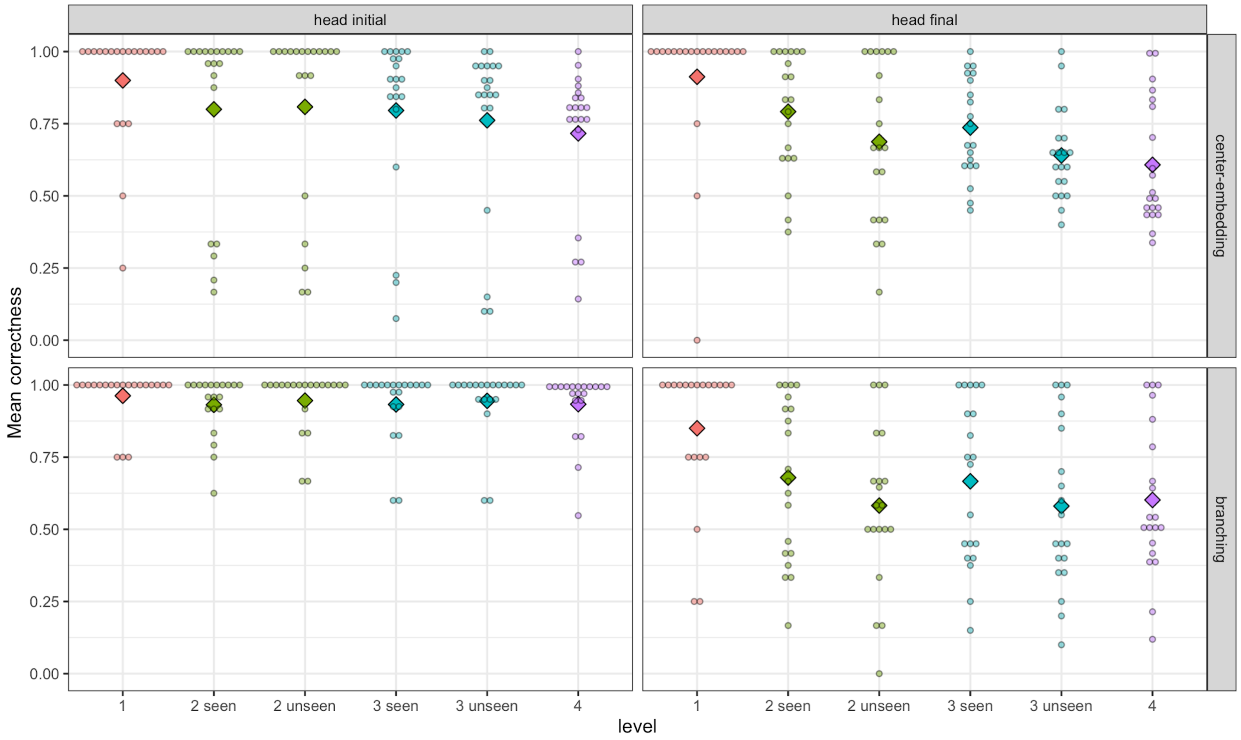

Figure 3 Write-in correctness for all arrays, pilot experiment 1

For write-in captions on arrays seen in training, performance was significantly worse in the head-final condition (B = -.057, SE = .023, p = .013). The effect of syntax type was not significant (B= 0.007, SE = 0.023, p = 0.757), but there was a significant interaction between conditions (B = -0.048, SE = 0.023 , p = 0.036). For write-in captions on novel arrays, similar patterns obtained. The effect of head order was significant (B = -.118, SE = .025, p < .001), while that of of syntax type was not (B= 0.031, SE = 0.025, p = 0.217), and there was a significant interaction between conditions (B = -0.059, SE = 0.025 , p = 0.019).

1.2.3. Grammars

Entropy was lower, i.e. the grammar was more consistent, in the branching condition (B= -1.057, SE = .36, p = .001), as seen in Figure 6. Figure 7 shows the grammar type for each string across conditions. Branching syntax was generally well-replicated, whereas center-embedded syntax was rarely replicated correctly, and sometimes converted to crossed syntax. Head order was often indeterminate due to unclear write-in adposition responses.

Summary table for binomial mixed effects model of comprehension task in Pilot Experiment 1.

Comprehension task

Row	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(level 1/branching/head-initial)	2.406	0.2	12.021	< .001
level2-1	0.526	0.315	1.671	0.095
level3-2	0.176	0.297	0.595	0.552
center-embedded	0.011	0.185	0.057	0.955
head-final	-0.392	0.185	-2.115	0.034
level2-1/center-embedded	0.073	0.238	0.305	0.76
level3-2/center-embedded	0.288	0.203	1.421	0.155
level2-1/head-final	-0.234	0.239	-0.98	0.327
level3-2/head-final	-0.303	0.207	-1.46	0.144
center-embedded/head-final	-0.161	0.185	-0.872	0.383
level2-1/center-embedded/head-final	0.145	0.237	0.613	0.54
level3-2/center-embedded/head-final	-0.333	0.202	-1.65	0.099

Summary table for mixed effects model of write-in task for Pilot Experiment 1. Level is coded as numeric.

Production task

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 1/branching/head-initial)	0.811	0.023	76.003	35.8	< .001
level	-0.05	0.006	76.003	-7.816	< .001
center-embedding	0.008	0.023	76.003	0.363	0.718
head-final	-0.072	0.023	76.003	-3.17	0.002
level/center-embedding	0.02	0.006	76.003	3.045	0.003
level/head-final	-0.022	0.006	76.003	-3.344	0.001
center-embedding/head-final	-0.052	0.023	76.003	-2.292	0.025
level/center-embedding/head-final	-0.003	0.006	76.003	-0.502	0.617

Summary table for mixed effects model of write-in performance on seen vs. novel arrays, Pilot Experiment 1.

Seen vs. unseen arrays

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 3 2/branching/head-initial/unseen)	0.799	0.033	79.052	24.122	< .001
level 3	-0.021	0.013	84.976	-1.67	0.099
center-embedding	-0.011	0.033	79.052	-0.33	0.742
head-final	-0.085	0.033	79.052	-2.559	0.012
unseen	-0.019	0.014	1746.513	-1.384	0.166
level 3/center-embedding	0.017	0.013	84.976	1.371	0.174
level 3/head-final	-0.008	0.013	84.976	-0.657	0.513
center-embedding/head-final	-0.065	0.033	79.052	-1.976	0.052
level 3/unseen	-0.003	0.009	1746.783	-0.364	0.716
center-embedding/unseen	-0.004	0.014	1746.513	-0.278	0.781
head-final/unseen	-0.036	0.014	1746.513	-2.598	0.009
level 3/center-embedding/head-final	0.005	0.013	84.976	0.369	0.713
level 3/center-embedding/unseen	0.006	0.009	1746.783	0.635	0.525
level 3/head-final/unseen	0.008	0.009	1746.783	0.897	0.37
center-embedding/head-final/unseen	0.005	0.014	1746.513	0.35	0.726
level 3/center-embedding/head-final/unseen	-0.005	0.009	1746.783	-0.537	0.591

Summary table for mixed effects model of write-in performance on seen arrays only, Pilot Experiment 1.

Production task: seen only

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 1/branching/head-initial)	0.83	0.023	75.926	36.863	< .001
level2-1	-0.106	0.024	76.024	-4.485	< .001
level3-2	-0.018	0.013	76.37	-1.347	0.182
center-embedding	0.007	0.023	75.926	0.311	0.757
head-final	-0.057	0.023	75.926	-2.536	0.013
level2-1/center-embedding	0.004	0.024	76.024	0.185	0.854
level3-2/center-embedding	0.012	0.013	76.37	0.921	0.36
level2-1/head-final	-0.04	0.024	76.024	-1.708	0.092
level3-2/head-final	-0.016	0.013	76.37	-1.167	0.247
center-embedding/head-final	-0.048	0.023	75.926	-2.138	0.036
level2-1/center-embedding/head-final	-0.03	0.024	76.024	-1.269	0.208
level3-2/center-embedding/head-final	0.009	0.013	76.37	0.707	0.482

Summary table for mixed effects model of write-in performance on unseen arrays only, Pilot Experiment 1.

Production task: novel only

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 2/branching/head-initial)	0.734	0.025	75.999	29.887	< .001
level3-2	-0.024	0.019	76.102	-1.293	0.2
level4-3	-0.017	0.013	76.196	-1.31	0.194
center-embedding	0.031	0.025	75.999	1.245	0.217
head-final	-0.118	0.025	75.999	-4.787	< .001
level3-2/center-embedding	0.023	0.019	76.102	1.221	0.226
level4-3/center-embedding	0.022	0.013	76.196	1.676	0.098
level3-2/head-final	0	0.019	76.102	-0.025	0.98
level4-3/head-final	0.011	0.013	76.196	0.88	0.382
center-embedding/head-final	-0.059	0.025	75.999	-2.404	0.019
level3-2/center-embedding/head-final	0	0.019	76.102	-0.003	0.997
level4-3/center-embedding/head-final	0.005	0.013	76.196	0.385	0.701

Summary table for linear regression of grammar distribution entropy by condition, Pilot Experiment 1.

Entropy by condition

Row	*Estimate*	*Std. Error*	*t value*	*Pr(>|t|)*
(center-embedding/head initial)	3.212	0.227	14.169	< .001
head final	0.158	0.321	0.493	0.623
branching	-1.057	0.321	-3.297	0.001
head final:branching	-0.095	0.453	-0.209	0.835

1.3. Discussion

Results for this experiment were broadly similar to those of the follow up experiment 1 (section 2 of main paper). Given the appearance of crossed grammar in the results as a best fit (and prior research such as Bach 1986 which indicates crossed dependencies may be cognitively easier than center-embedding), we used crossed grammar as an additional syntactic condition in Experiment 1. In addition, adposition training was included to help clarify the resulting participant grammars, which could not always be disambiguated due to unclear adposition meanings.

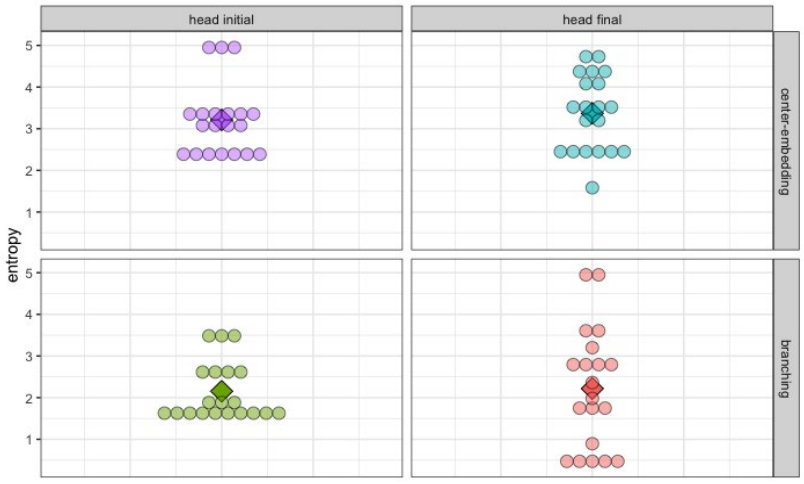

Figure 6: entropy of participant grammar distributions by condition

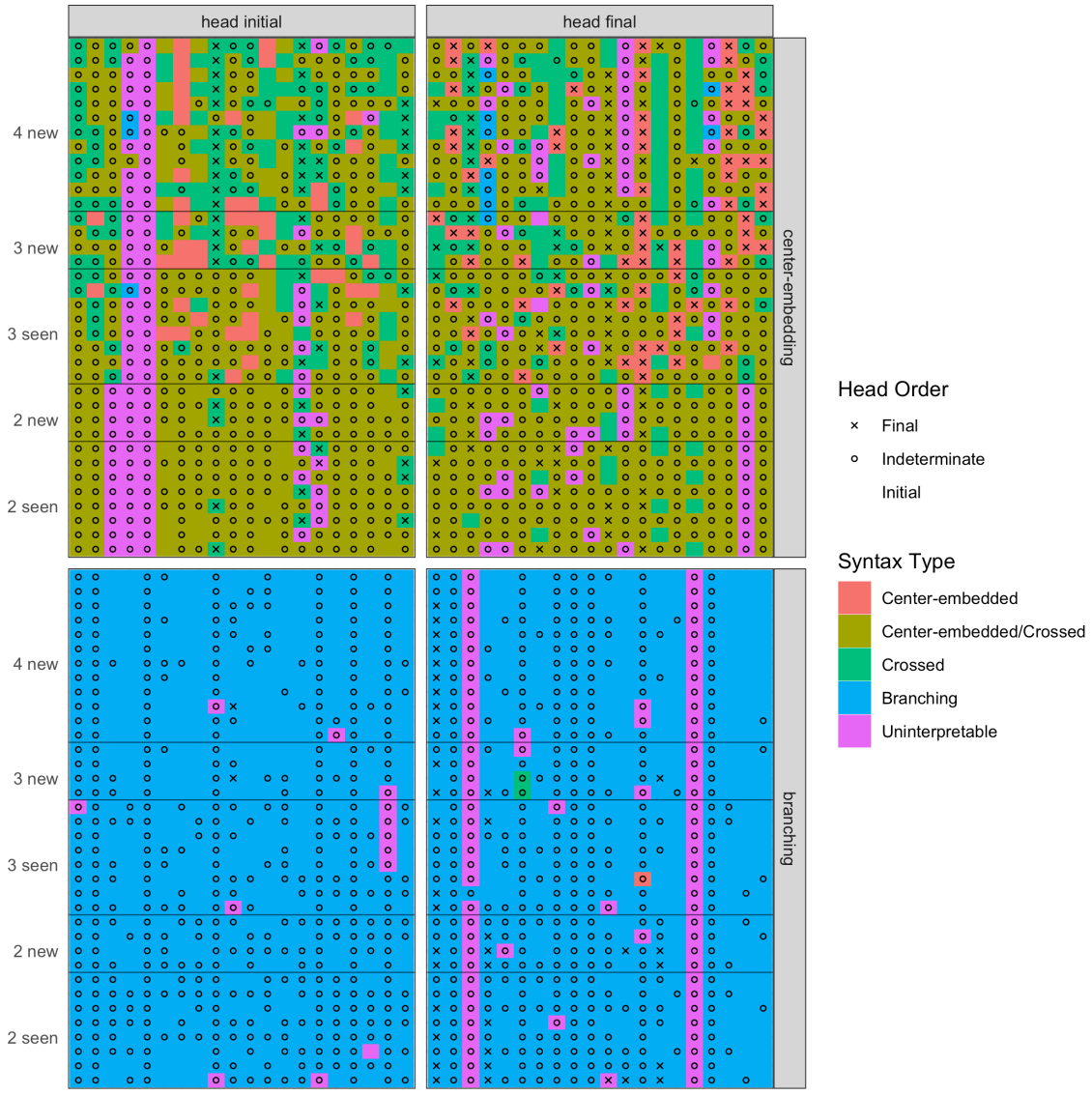

Figure 7: grammar by string, pilot experiment 1.

2. Pilot Experiment 2: Emergence of locative order in iterated learning

2.1. Methodology

2.1.1. Participants

As in the previous experiment, participants were recruited online through Mechanical Turk and compensated $5 for their time. Data used for analysis came from 100 total participants, forming ten chains of ten generations each. All were native English speakers according to self-report (in the post-task survey), though some spoke other languages.

2.1.2. Stimuli

Stimuli and lexica of the artificial languages were the same as those used previously. Each chain was assigned a randomly chosen set of 4 nouns from the 10 sets.

Each participant's language consisted of 40 arrays (same for all participants), each with a label. The first 4 of the arrays are individual objects, 12 each of 2, 3, and 4 objects. At generation 0, this language was randomly generated (see below); at each subsequent

generation, the previous participant's labels from the writing task are used for training. Random stimuli for generation 0 were produced as in experiment 2.

2.1.3. Procedure

The procedure was as in experiment 2, but without adposition training or training to criterion, and no visual component in the final survey.

2.1.4. Analysis

Analysis was the same as experiment 2, except that level 1 (1-item data) was included in comprehension results as it was not trained to criterion.

2.2. Results

Complete statistical analysis of all measures can be found in Appendix 2, section 2.

2.2.1. Comprehension

Comprehension scores (Figure 8) showed a slight but not significant increase over the generations (B = .072, SE = .084, p = .393). The overall effect of level, however, was significant from level 1 to level 2 (B = -1.313, SE = .531, p =.013) but not level 2 to 3 (B = -.463, SE = .306, p =.393).

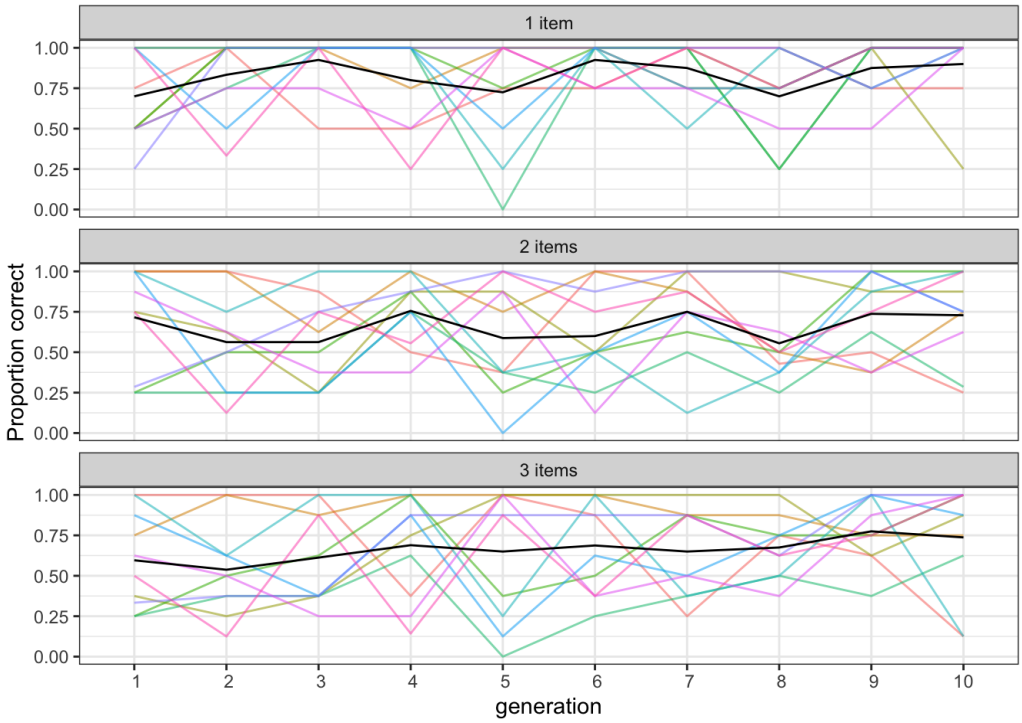

Figure 8. Comprehension accuracy by generation and number of objects, Pilot Experiment 2

2.2.2. Production

Figure 9 shows accuracy and consistency of productions. There was a general uptrend in the accuracy of productions for arrays that were seen (B = 0.032, SE = 0.013, 0= 0.034), and a non-significant increase in the consistency of descriptions for novel arrays (B = 0.025, SE = 0.016, p = 0.146).

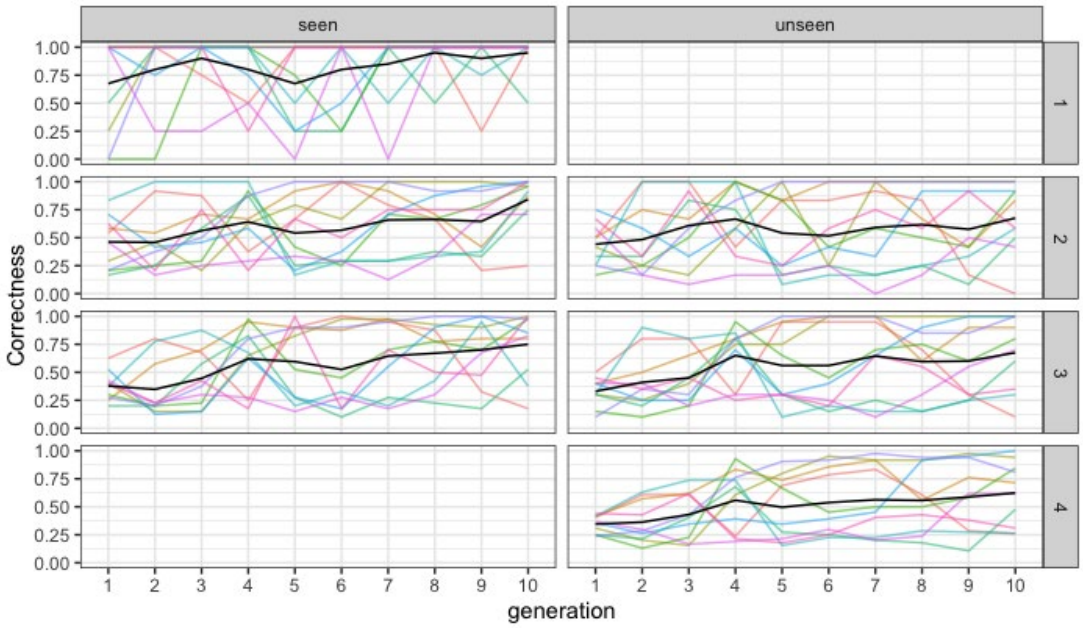

Figure 9: Production accuracy for seen captions, and consistency of captions for new arrays

2.2.3. Grammars

Entropy of grammar distributions (figure 10) decreased significantly over the generations (B= -.117, SE = .03, p < .001).

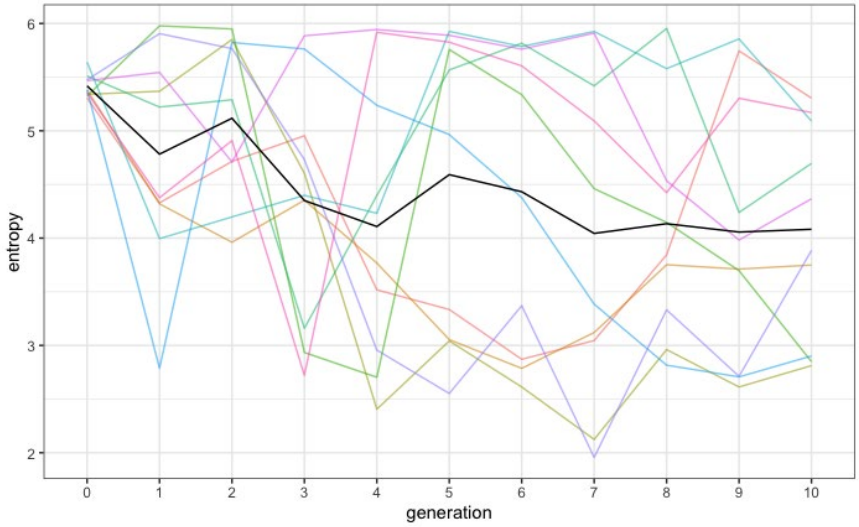

Figure 10: entropy of grammar distribution by generation and chain

Figure 11 shows the emergence of stable grammars over the generations for each chain. As can be seen, branching grammar came to dominate in most chains.

2.3. Discussion

The results of this study are consistent with the previous learnability results and with psycholinguistic research concerning the difficulty of center-embedding; broadly similar results also occurred in the follow-up experiment (section 3 of main paper). Results were chaotic in

some chains (see chains E, F, and I in figure 11), but others clearly stabilized on a consistent branching grammar after 3-5 generations.

As in the first pilot experiment, the lack of clearly defined adposition interpretations made it difficult to narrow down grammars-- and specifically to determine whether participants were following English syntax (branching head-initial), or producing branching grammar with either head order. Consequently, adposition training was incorporated into future experiments, and for the iterated learning experiment (section 3 of main paper), adposition meaning was iterated along with array captions.

Summary table for binomial mixed effects model of comprehension task in Pilot Experiment 2.

Comprehension task

Row	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
(level 1/generation 1)	1.003	0.53	1.892	0.058
level2-1	-1.323	0.531	-2.492	0.013
level3-2	-0.463	0.306	-1.512	0.13
generation	0.072	0.084	0.854	0.393
level2-1/generation	-0.039	0.081	-0.484	0.628
level3-2/generation	0.086	0.049	1.735	0.083

Summary table for mixed effects model of caption consistency for captions seen in training, Pilot Experiment 2.

Consistency of captions: arrays seen in training

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 1/generation 0)	0.49	0.076	9.001	6.423	< .001
level2-1	-0.279	0.077	22.403	-3.63	0.001
level3-2	-0.092	0.076	58.144	-1.211	0.231
generation	0.032	0.013	9.006	2.505	0.034
level2-1/generation	0.01	0.013	19.218	0.746	0.465
level3-2/generation	0.01	0.012	48.812	0.813	0.42

Summary table for mixed effects model of caption consistency for captions on novel arrays, Pilot Experiment 2.

Consistency of captions: novel arrays

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(level 2/generation 1)	0.404	0.072	8.998	5.623	< .001
level3-2	-0.111	0.065	157.967	-1.714	0.088
level4-3	-0.029	0.066	78.407	-0.433	0.666
generation	0.025	0.016	8.997	1.59	0.146
level3-2/generation	0.016	0.011	62.458	1.493	0.14
level4-3/generation	-0.002	0.011	80.736	-0.215	0.831

Summary table of linear regression model of grammar distribution entropy over generations, Pilot Experiment 2.

Entropy

Row	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Generation 0)	5.164	0.263	29.861	19.653	< .001
Generation	-0.117	0.03	99	-3.883	< .001

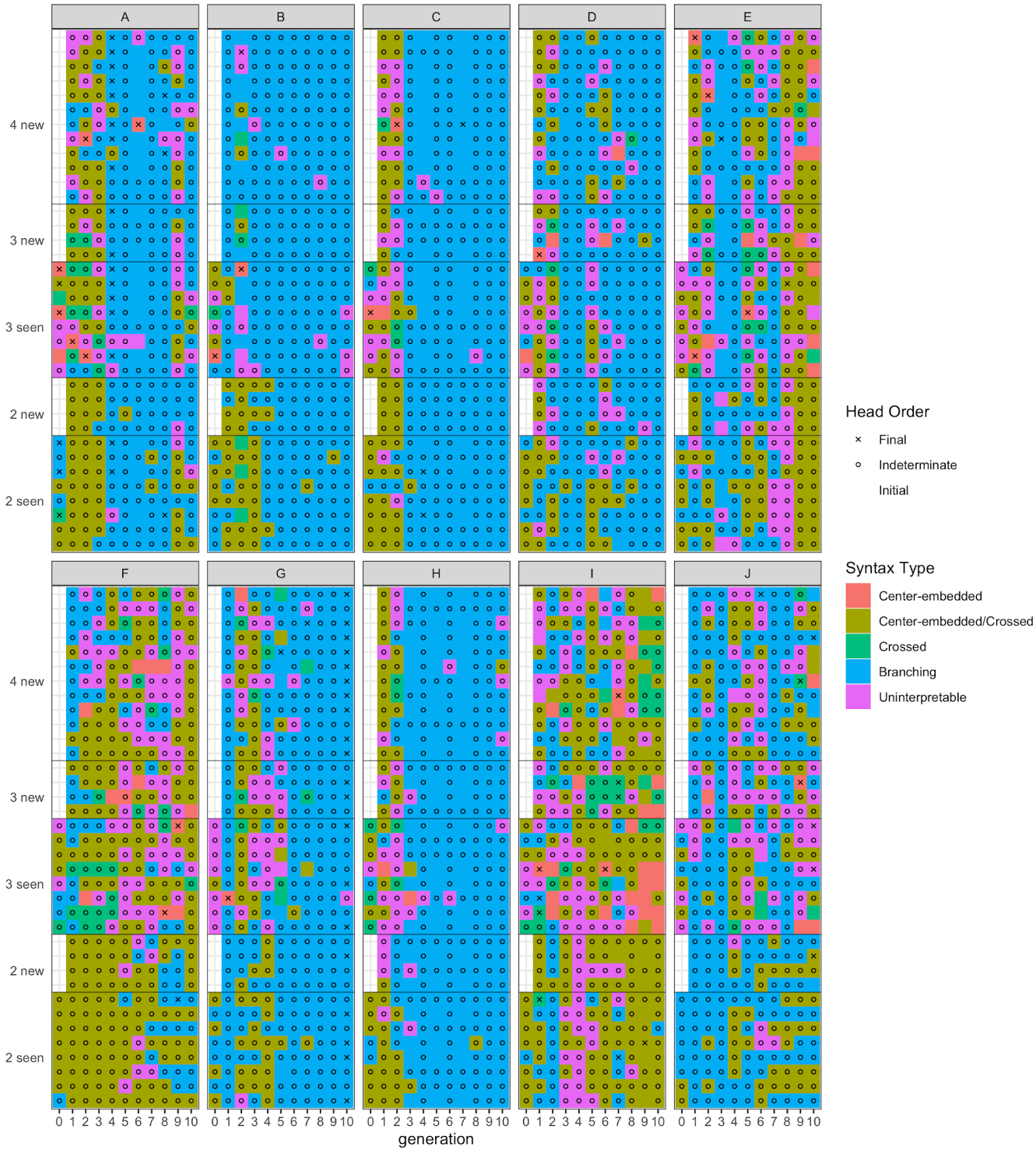

Figure 11: grammar by string, chain, and generation, pilot experiment 2.