

Larger communities create more systematic languages

Limor Raviv^{1,a}, Antje Meyer^{a,b} & Shiri Lev-Ari^{a,c}

Keywords: language evolution; linguistic diversity; grammatical structure; social structure; community size

¹ Corresponding author: limor.raviv@mpi.nl +31-624751126
ORCID: <https://orcid.org/0000-0002-0716-3553>

^a Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

^b Radboud University, Comeniuslaan 4, 6525 HP Nijmegen, The Netherlands; antje.meyer@mpi.nl

^c Royal Holloway University of London, Egham Hill, Egham TW20 0EX, UK; shiri.levvari@rhul.ac.uk

Author Contribution: L.R. designed and performed the research, analyzed the data, and wrote the paper; A.M. and S.L. designed the research and wrote the paper.

Abstract

Understanding world-wide patterns of language diversity has long been a goal for evolutionary scientists, linguists and philosophers. Research over the past decade suggested that linguistic diversity may result from differences in the social environments in which languages evolve. Specifically, recent work found that languages spoken in larger communities typically have more systematic grammatical structures. However, in the real world, community size is confounded with other social factors such as network structure and the number of second languages learners in the community, and it is often assumed that linguistic simplification is driven by these factors instead. Here we show that in contrast to previous assumptions, community size has a unique and important influence on linguistic structure. We experimentally examine the live formation of new languages created in the lab by small and larger groups, and find that larger groups of interacting participants develop more systematic languages over time, and do so faster and more consistently than small groups. Small groups also vary more in their linguistic behaviors, suggesting that small communities are more vulnerable to drift. These results show that community size predicts patterns of language diversity, and suggest that an increase in community size might have contributed to language evolution.

Introduction

39
40 Almost 7,000 languages are spoken around the world (1,2), and the remarkable range of linguistic diversity
41 has been studied extensively (3,4). Current research focuses on understanding the sources for this
42 diversity, and attempts to understand whether differences between languages can be predicted by
43 differences in their environments (5–11). If languages evolved as a means for social coordination (12,13),
44 they are bound to be shaped by their social environment and the properties of the cultures in which they
45 evolved. Indeed, cross-linguistic and historical studies have suggested that different linguistic structures
46 emerge in different societies depending on their size, network structure, and the identity of their members
47 (5,14-18).

48 One social property, community size, might play a particularly important role in explaining
49 grammatical differences between languages. First, an increase in human group size was argued to be one
50 of the drivers for the evolution of natural language (19). Second, cross-linguistic work that examined
51 thousands of languages found that languages spoken in larger communities tend to be less complex (5).
52 Specifically, these languages have fewer and less elaborate morphological structures, fewer irregulars,
53 and overall simpler grammars (5). In addition to shaping grammar, community size could affect trends of
54 convergence and stability during language change (14-18).

55 While there is correlational evidence for the relation between community size and grammatical
56 complexity, cross-linguistic studies cannot establish a causal link between them. Furthermore, the
57 relationship between bigger communities and linguistic simplification can be attributed to other social
58 factors that are confounded with community size in the real world. In particular, bigger communities tend
59 to be more sparsely connected, more geographically spread out, have more contact with outsiders, and
60 have a higher proportion of adult second language learners (14-16). Each of these factors may contribute
61 to the pattern of reduced complexity, and thus provide an alternative explanation for the correlation
62 between community size and linguistic structure (5-8,20-21). In fact, many researchers assume that this
63 correlation is accounted for by the proportion of second language learners in the community (5-7,20) or
64 by differences in network connectivity (15-17,21; See discussion).

65 Here we argue that community size has a unique and casual role in explaining linguistic diversity, and
66 show that it influences the formation of different linguistic structures in the evolution of new languages.
67 Interacting with more people reduces shared history and introduces more input variability (i.e., more
68 variants), which individuals need to overcome before the community can reach mutual understanding.
69 Therefore, interacting with more people can favor systematization by introducing a stronger pressure for
70 generalizations and transparency. That is, larger communities may be more likely to favor linguistic
71 variants that are simple, predictable, and structured, which can in turn ease the challenge of convergence
72 and communicative success. Supporting this idea, language learning studies show that an increase in input
73 variability (i.e., exposure to multiple speakers) boosts categorization, generalization, and pattern detection
74 in infants and adults (22–29).

75 While existing studies cannot establish a causal link between community structure and linguistic
76 structure or isolate the role of community size, teasing apart these different social factors has important
77 implications for our understanding of linguistic diversity and its origins (30). Some computational models
78 attempted to isolate the effect of community size on emerging languages using populations of interacting
79 agents, but their results show a mixed pattern: while some models suggest that population size plays little
80 to no role in explaining cross-linguistic patterns (21,31,32), others report strong associations between
81 population size and linguistic features (33-35).

82 To date, no experimental work has examined the effect of community size on the emergence of
83 language structure with human participants, although it was suggested several times (36–38). We fill this
84 gap by conducting a behavioral study that examines the live formation of new communicative systems
85 created in the lab by small or larger groups. A couple of previous studies investigated the role of input
86 variability, one of our hypothesized mechanisms, using an individual learning task, yet found no effect of

87 learning from different models (39,40). Another related study compared the complexity of English
88 descriptions produced for novel icons by two or three people, but reported no differences between the
89 final descriptions of dyads and triads (41). These studies, however, did not test the emergence of
90 systematic linguistic structure. Here we examine how group size influences the emergence of
91 compositionality in a new language, and assess the role of input variability in driving this effect. In
92 addition to examining changes in linguistic structure over time, we track other important aspects of the
93 emerging systems (e.g., communicative success and the degree to which languages are shared across
94 participants), shedding light on how community size affects the nature of emerging languages.

95 **The Current Study**

96 We used a group communication paradigm inspired by (42-47) to examine the performance of small
97 and larger microsocieties. Participants interacted in alternating pairs with the goal of communicating
98 successfully using only an artificial language they invented during the experiment. In each communication
99 round, paired partners took turns in describing novel scenes of moving shapes, such that one participant
100 produced a label to describe a target scene, and their partner guessed which scene they meant from a larger
101 set of scenes. Participants in small and larger groups had the same amount of interaction overall, but
102 members of larger groups had less shared history with each other by the end of the experiment. All other
103 group properties (e.g., network structure) were kept constant across conditions.

104 We examined the emerging languages over the course of the experiment using several measurements
105 (see Measures): (1) Communicative Success; (2) Convergence, reflecting the degree of alignment in the
106 group (3) Stability, reflecting the degree of change over time; and (4) Linguistic Structure, reflecting the
107 degree of systematic mappings in the language. With these measures, we can characterize the emerging
108 communication systems and understand how different linguistic properties change over time depending
109 on community size.

110 Our main prediction was that larger groups would create more structured languages, given that they
111 are under a stronger pressure for generalization due to increased input variability and reduced shared
112 history. We also predicted that larger groups would show slower rates of stabilization and convergence
113 compared to smaller groups. Furthermore, we ran analyses to test our proposed mechanism, namely, that
114 larger groups create more structured languages because of greater input variability and reduced shared
115 history.

116 **Methods**

117 **Participants**

118 Data from 144 adults (mean age=24.9y, SD=8.9y; 103 women) was collected over the period of one year
119 in several batches, comprising 12 small groups of four members and 12 larger groups of eight members.
120 Participants were paid 40€ or more depending on the time they spent in the lab (between 270 to 315
121 minutes, including a 30-minutes break). Six additional small groups took part in a shorter version of the
122 experiment (47), which included only eight rounds. These additional groups showed similar patterns of
123 results when compared to the larger groups. Their results are reported in Appendix B. All participants
124 were native Dutch speakers. Ethical approval was granted by the Faculty of Social Sciences of the
125 Radboud University Nijmegen.

126
127
128
129
130
131
132
133
134
135
136
137

Materials

We created visual scenes that varied along three semantic dimensions: shape, angle of motion, and fill pattern (see also 44,45,47). Each scene included one of four novel shapes, moving repeatedly in a straight line from the center of the frame in an angle chosen from a range of possible angles. The four shapes were unfamiliar and ambiguous in order to discourage labeling with existing words. Angle of motion was a continuous feature, which participants could have parsed and categorized in various ways. Additionally, the shape in each scene had a unique blue-hued fill pattern, giving scenes an idiosyncratic feature. Therefore, the meaning space promoted categorization and structure along the dimensions of shape and motion, but also allowed participants to adopt a holistic, unstructured strategy where scenes are individualized according to their fill pattern. There were three versions of the stimuli, which differed in the distribution of shapes and their associated angles (see Appendix A). Each version contained 23 scenes and was presented to two groups in each condition. The experiment was programmed using Presentation.

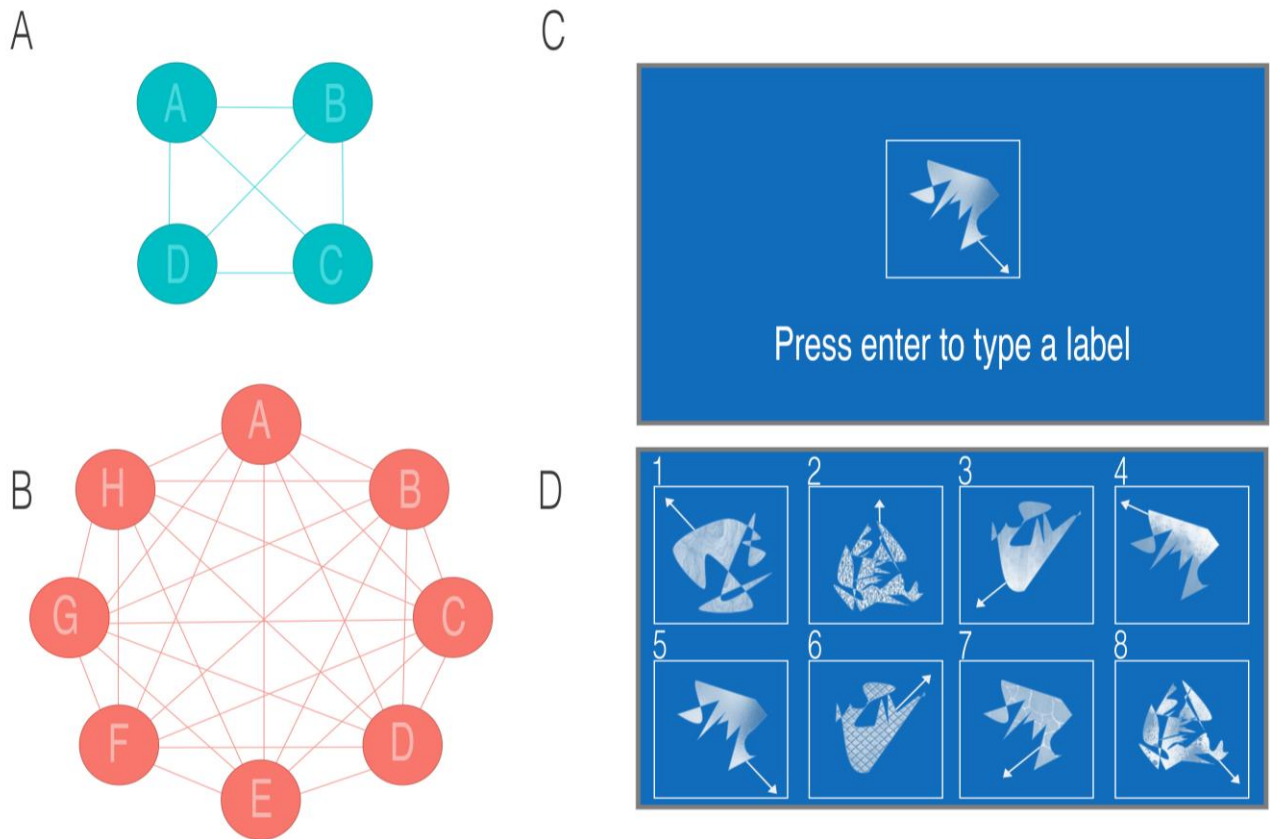


Figure 1. Group communication paradigm. We tested fully-connected groups of either four (A) or eight (B) participants. Panels (C) and (D) show the producer's and guesser's screens, respectively.

138
139
140
141
142
143
144

Procedure

Participants were asked to create a fantasy language and use it in order to communicate about different novel scenes. Participants were not allowed to communicate in any other way besides typing, and their letter inventory was restricted: it included a hyphen, five vowel characters (a,e,i,o,u) and ten consonants (w,t,p,s,f,g,h,k,n,m), which participants could combine freely.

The experiment had 16 rounds, comprising three phases: group naming (round 0), communication (rounds 1-7; rounds 9-15), and test (round 8; round 16).

145 In the naming phase (round 0), participants generated novel nonsense words to describe eight initial
146 scenes, so that each group had a few shared descriptions to start with. Eight scenes were randomly drawn
147 from the set of 23 scenes (see Materials) under the constraint that each shape and quadrant were
148 represented at least once. During this phase, participants sat together and took turns in describing the
149 scenes, which appeared on a computer screen one by one in a random order. Participants in larger groups
150 named one scene each, and participants in small groups naming two scenes each. Importantly, no use of
151 Dutch or any other language was allowed. An experimenter was present in the room throughout the
152 experiment to ensure participants did not include known words. Once a participant had typed a description
153 for a scene, it was presented to all group members for several seconds. This procedure was repeated until
154 all scenes had been named and presented once. In order to establish shared knowledge, these scene-
155 description pairings were presented to the group twice more in a random order.

156 Following the naming phase, participants played a communication game (the communication phase):
157 the goal was to earn as many points as possible as a group, with a point awarded for every successful
158 interaction. The experimenter stressed that this was not a memory game, and that participants were free
159 to use the labels produced during the group naming phase, or create new ones. Paired participants sat on
160 opposite sides of a table facing each other and personal laptop screens (see Appendix A). During this
161 phase, group members exchanged partners at the start of every round, such that by end of the experiment,
162 each pair in the small group has interacted at least four times and each pair in the large group has interacted
163 exactly twice.

164 In each communication round, paired participants interacted 23 times, alternating between the roles of
165 producer and guesser. In each interaction, the producer saw the target scene on their screen (see Fig. 1C)
166 and typed a description using their keyboard. The guesser saw a grid of eight scenes on their screen (the
167 target and seven distractors), and had to press the number associated with the scene they thought their
168 partner referred to. Participants then received feedback on their performance.

169 The number of target scenes increased gradually over the first six rounds, such that participants
170 referred to more scenes in later rounds. While round 1 included only the eight initial scenes selected for
171 the group naming phase, three new scenes were added in each following round until there were 23 different
172 scenes in round 6. No more scenes were introduced afterwards, allowing participants to interact about all
173 scenes for the following rounds. This method was implemented in order to introduce a pressure for
174 developing structured and predictable languages (47), and resembles the real world with its unconstrained
175 meaning space.

176 After the seventh communication round, participants completed an individual test phase (round 8), in
177 which they typed their descriptions for all scenes one by one in a random order. After the test, participants
178 had seven additional communication rounds (rounds 9-15) and the additional test round (round 16). These
179 two individual test rounds allowed us to get a full representation of participants' entire lexicon at the
180 middle and end of the experiment. Finally, participants filled out a questionnaire about their performance
181 and were debriefed by the experimenter.

182 Due to a technical error, one large group played only six additional communication rounds instead of
183 seven. Additionally, data from one participant in a large group was lost. The existing data from these
184 groups was included in the analyses.

185 **Measures**

186 *Communicative Success*

187 Measured as binary response accuracy in a given interaction during the communication phase, reflecting
188 comprehension.

189 *Convergence*

190 Measured as the similarities between all the labels produced by participants in the same group for the same
191 scene in a given round: for each scene in round n , convergence was calculated by averaging over the
192 normalized Levenshtein distances between all labels produced for that scene in that round. The normalized
193 Levenshtein distance between two strings is the minimal number of insertions, substitutions, and deletions
194 of a single character that is required for turning one string into the other, divided by the number of
195 characters in the longer string. This distance was subtracted from 1 to represent string similarity, reflecting
196 the degree of shared lexicon and alignment in the group.

197 *Stability*

198 Measured as the similarities between the labels created by participants for the same scenes on two
199 consecutive rounds: for each scene in round n , stability was calculated by averaging over the normalized
200 Levenshtein distances between all labels produced for that scene in round n and round $n+1$. This distance
201 was subtracted from 1 to represent string similarity, reflecting the degree of consistency in the groups'
202 languages.

203 *Linguistic Structure*

204 Measured as the correlations between string distances and semantic distances in each participant's
205 language in a given round, reflecting the degree to which similar meanings are expressed using similar
206 strings (43,44,47). First, scenes had a semantic difference score of 1 if they differed in shape, and 0
207 otherwise. Second, we calculated the absolute difference between scenes' angles, and divided it by the
208 maximal distance between angles (180 degrees) to yield a continuous normalized score between 0 and 1.
209 Then, the difference scores for shape and angle were added, yielding a range of semantic distances
210 between 0.18 and 2. Finally, labels' string distances were calculated using the normalized Levenshtein
211 distances between all possible pairs of labels produced by participant p for all scenes in round n . For each
212 participant, the two sets of pair-wise distances (i.e., string distances and meaning distances) were
213 correlated using the Pearson product-moment correlation. While most iterated learning studies use the z-
214 scores provided by the Mantel test for the correlation described above (43,44), z-scores were inappropriate
215 for our design since they increase with the number of observations, and our meaning space expanded over
216 rounds. Therefore, we used the raw correlations between meanings and strings as a more accurate measure
217 of systematic structure (47, 48).

218 *Input Variability*

219 Measured as the minimal sum of differences between all the labels produced for the same scene in a given
220 round. For each scene in round n , we made a list of all label variants for that scene. For each label variant,
221 we summed over the normalized Levenshtein distances between that variant and all other variants in the
222 list. We then selected the variant that was associated with the lowest sum of differences (i.e., the 'typical'
223 label), and used that sum as the input variability score for that scene, capturing the number of different
224 variants and their relative difference from each other. Finally, we averaged over the input variability scores
225 of different scenes to yield the mean variability in that round.

226 *Shared History*

227 Measured as the number of times each pair in the group interacted so far, reflecting the fact that small
228 groups interacted more often with each other. In small groups, pairs interacted once by round 3, twice by
229 round 6, three times by round 10, four times by round 14, and started to interact for the fifth time in round
230 15. In larger groups, pairs only interacted once by round 7, and twice by round 15.

231 **Analyses**

232 We used mixed-effects regression models to test the effect of community size on all measures using the
233 lme4 (49) and pbkrtest (50) packages in R (51). All models had the maximal random effects structure
234 justified by the data that would converge. The reported p-values were generated using the Kenward-Roger
235 Approximation, which gives more conservative p-values for models based on small numbers of
236 observations. The full models are included in Appendix C. All the data and the scripts for generating all
237 models can be openly found at <https://osf.io/y7d6m/>.

238 Changes in communicative success, stability, convergence and linguistic structure were examined
239 using three types of models: (I) Models that analyze changes in the dependent variable over time; (II)
240 Models that compare the final levels of the dependent variable at the end of the experiment; (III) Models
241 that examine differences in the levels of variance in the dependent variable over time.

242 Models of type (I) predicted changes in the dependent variable as a function of time and community
243 size. Models for communicative success included data from communication rounds only (excluding the
244 two test rounds). In models for communicative success, convergence, and stability, the fixed effects were
245 CONDITION (dummy-coded with small group as the reference level), ROUND NUMBER (centered), ITEM
246 CURRENT AGE (centered), and the interaction terms CONDITION X ITEM CURRENT AGE and CONDITION X
247 ROUND NUMBER. ITEM CURRENT AGE codes the number of rounds each scene was presented until that point
248 in time, and measures the effect of familiarity with a specific scene on performance. ROUND NUMBER
249 measures the effect of time passed in the experiment and overall language proficiency. The random effects
250 structure of models for communicative success, convergence, and stability included by-scenes and by-
251 groups random intercepts, as well as by-groups random slopes for the effect of ROUND NUMBER. Models
252 from stability and communicative success also included by-scenes random slopes for the effect of ROUND
253 NUMBER. As structure score was calculated for each producer over all scenes in a given round, the model
254 for linguistic structure did not include ITEM CURRENT AGE as a fixed effect, and included fixed effects for
255 ROUND NUMBER (quadratic, centered), CONDITION (dummy-coded with small group as the reference level),
256 and the interaction term CONDITION X ROUND NUMBER. Following Beckner et al. (2017)(52), who found
257 that linguistic structure tends to increase nonlinearly, we included both the linear and the quadratic terms
258 (using the poly() function in R to avoid colinearity). The model for linguistic structure included random
259 intercepts and random slopes for the effect of ROUND NUMBER with respect to different producers who
260 were nested in different groups.

261 Models of type (II) compared the mean values of the final languages created by small and larger groups
262 in rounds 15-16. The fixed effect in these models was a two-level categorical variable (i.e., small groups
263 vs. larger groups), dummy-coded with small groups as the reference level. In models for communicative
264 success, stability and structure, the random effects structure included random intercepts for different groups
265 and different scenes. In models for linguistic structure, the random effect structure included random
266 intercepts for different producers nested in different groups.

267 Models of type (III) predicted the degree of variance in the dependent variable across groups and time.
268 For linguistic structure, variance was calculated as the square standard deviation in participants' average
269 structure scores across all groups in a given round. For communicative success, convergence and stability,
270 variance was calculated as the square standard deviation in the dependent variable on each scene across
271 all groups in a given round. These models included by-scenes random intercepts and slopes for the effect

272 of ROUND NUMBER. All models included fixed effects for ROUND NUMBER (centered), CONDITION (dummy-
273 coded with small group as the reference level), and the interaction term CONDITION X ROUND NUMBER.

274 We also examined changes in input variability as a function of time and community size. This model
275 included fixed effects for ROUND NUMBER (centered), CONDITION (dummy-coded with small group as the
276 reference level), and the interaction between them. There was a by-group random intercepts and by-group
277 random slopes for the effect of ROUND NUMBER. Finally, we examined changes in linguistic structure
278 scores over consecutive rounds as a function of (a) input variability, (b) shared history, or (c) both. In all
279 three models, the dependent variable was the difference in structure score between round n and $n+1$, and
280 there were random intercepts for different producers nested in different groups. In model (a), the fixed
281 effect was MEAN INPUT VARIABILITY at round n (centered). In model (b), the fixed effect was SHARED
282 HISTORY at round n (centered). Model (c) was a combination of models (a) and (b).

283

Results

284 We report the results for each of the four linguistic measures separately, using three types of analyses (see
285 Methods). Figure 2 summarizes the average differences in the performance of small and larger groups
286 over the course of all 16 rounds. Note that all analyses were carried over all data points and not over
287 averages. All analyses are reported in full in Appendix C using numbered models, which we refer to here.

288

1. Communicative Success

289 Communicative Success increased over time (Model 1: $\beta=0.08$, $SE=0.02$, $t=4$, $p<0.0001$; Fig. 2A), with
290 participants becoming more accurate as rounds progressed. This increase was not significantly modulated
291 by group size (Model 1: $\beta=0.04$, $SE=0.03$, $t=1.76$, $p=0.078$), with small and larger groups reaching similar
292 accuracy scores in the final communication round (Model 2: $\beta=0.14$, $SE=0.08$, $t=1.8$, $p=0.083$). Small and
293 larger groups differed in variance: while all groups became increasingly more varied over time (Model 3:
294 $\beta=0.002$, $SE=0.0004$, $t=5.18$, $p<0.0001$), larger groups showed a slower increase in variance (Model 3:
295 $\beta=-0.002$, $SE=0.0005$, $t=-4.2$, $p<0.0001$) and lower variance overall (Model 3: $\beta=-0.007$, $SE=0.002$, $t=-$
296 3.48 , $p<0.001$). These results indicate that while small groups varied in their achieved accuracy scores,
297 and even more so as the experiment progressed, larger groups tended to behave more similarly to one
298 another throughout the experiment.

299

2. Convergence

300 Convergence increased significantly across rounds (Model 4: $\beta=0.007$, $SE=0.003$, $t=2.31$, $p=0.029$; Fig.
301 2B), with participants aligning and using more similar labels over time. Convergence was also better on
302 more familiar scenes (Model 4: $\beta=0.004$, $SE=0.001$, $t=2.62$, $p=0.014$). Group size had no effect on
303 convergence (Model 4: $\beta=-0.06$, $SE=0.04$, $t=-1.37$, $p=0.18$), so that small and larger groups showed similar
304 levels of convergence by the end of the experiment (Model 5: $\beta=-0.03$, $SE=0.05$, $t=-0.63$, $p=0.54$).
305 Interestingly, larger groups were not less converged than small groups, despite the fact that members of
306 larger groups had double the amount of people to converge with and only half the amount of shared history
307 with each of them. Variance increased over rounds (Model 6: $\beta=0.001$, $SE=0.003$, $t=4.32$, $p<0.0001$), but
308 there was significantly less variance in the convergence levels of larger groups than across small groups
309 throughout the experiment (Model 6: $\beta=-0.04$, $SE=0.002$, $t=-23.68$, $p<0.0001$). That is, larger groups
310 behaved similarly to each other, showing a slow yet steady increase in convergence over rounds, while
311 small groups varied more in their behavior: some small groups reached high levels of convergence, but
312 others maintained a high level of divergence throughout the experiment, with different participants using
313 their own unique labels.

314 3. Stability

315 Stability significantly increased over time, with participants using labels more consistently as rounds
316 progressed (Model 7: $\beta=0.009$, $SE=0.003$, $t=3.26$, $p=0.003$; Fig. 2C). Labels for more familiar scenes were
317 also more stable (Model 7: $\beta=0.004$, $SE=0.001$, $t=3.68$, $p=0.001$). Group size affected stability (Model 7:
318 $\beta=-0.08$, $SE=0.04$, $t=-2.08$, $p=0.047$), with larger groups' languages being less stable (i.e., showing more
319 changes). However, by the end of the experiment, the languages of small and larger groups did not differ
320 in their stability (Model 8: $\beta=-0.06$ $SE=0.05$, $t=-1.21$, $p=0.24$). As in the case of convergence, larger
321 groups showed significantly less variance in their levels of stability compared to small groups throughout
322 the experiment (Model 9: $\beta=-0.018$, $SE=0.001$, $t=-16.99$, $p<0.0001$), reflecting the fact that smaller groups
323 differed more from each other in their stabilization trends.

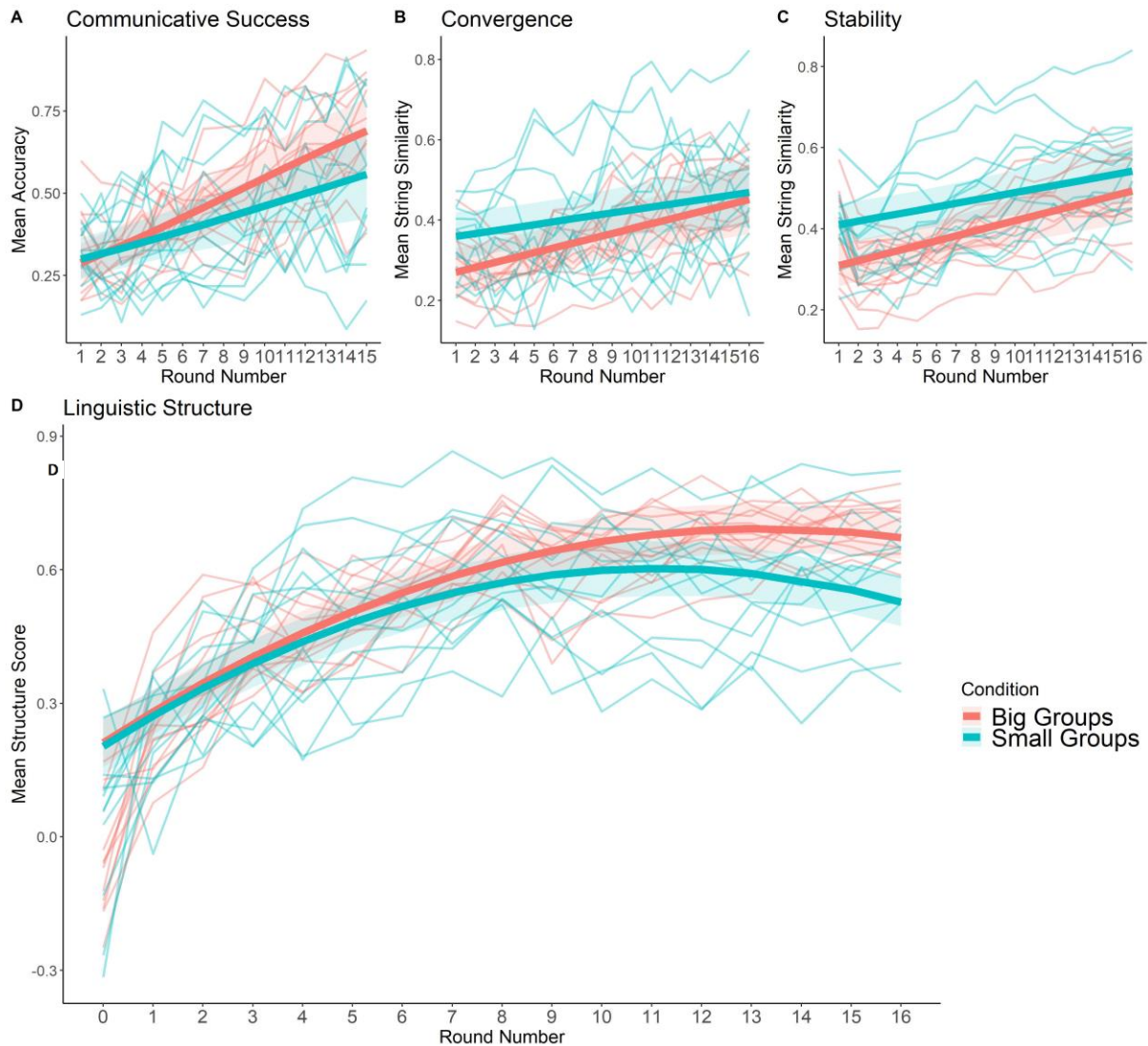


Figure 2. Changes in (A) Communicative Success, (B) Convergence, (C) Stability, and (D) Linguistic Structure over time as a function of community size. Thin lines represent average values for each group in a given round. Data from small and larger groups is plotted in blue and red, respectively. Thick lines represent the models' estimates, and their shadings represent the models' standard errors.

324 4. Linguistic Structure

325 Linguistic Structure significantly increased over rounds (Model 10: $\beta=4.55$, $SE=0.48$, $t=9.46$, $p<0.0001$;
326 Fig 2D), with participants' languages becoming more systematic over time. This increase was non-linear
327 and slowed down in later rounds (Model 10: $\beta=-3$, $SE=0.38$, $t=-7.98$, $p<0.0001$). As predicted, the increase
328 in structure was significantly modulated by group size (Model 10: $\beta=1.92$, $SE=0.63$, $t=3.06$, $p=0.004$), so
329 that participants in larger groups developed structured languages faster compared to participants in small
330 groups. Indeed, the final languages developed in larger groups were significantly more structured than the
331 final languages developed in small groups (Model 11: $\beta=0.11$, $SE=0.04$, $t=2.93$, $p=0.006$). Variance did
332 not significantly decrease over time (Model 12: $\beta=-0.0009$, $SE=0.0005$, $t=-1.73$, $p=0.094$), yet larger
333 groups varied significantly less overall in how structured their languages were (Model 12: $\beta=-0.015$,
334 $SE=0.004$, $t=-4.28$, $p=0.0002$). That is, while small groups differed in their achieved levels of structure
335 throughout the experiment, different larger groups showed similar trends and reached similar structure
336 scores.

337 Although all groups started out with different random holistic labels, compositional languages emerged
338 in many groups during the experiment. Many groups developed languages with systematic and predictable
339 grammars (see Fig. 3 for one example, and Appendix D for more examples), in which scenes were
340 described using complex labels: one part indicating the shape, and another part indicating motion¹.
341 Interestingly, groups differed not only in their lexicons, but also in the grammatical structures they used
342 to categorize scenes according to motion. While many groups categorized angles based on a two axes
343 system (with part-labels combined to indicate up/down and right/left), other groups parsed angles in a
344 clock-like system, using unique part-labels to describe different directions. Importantly, while no two
345 languages were identical, the level of systematicity in the achieved structure depended on group size.

346 We also tested our hypothesis that group size effects are driven by differences in input variability and
347 shared history. First, we quantified the degree of input variability in each group at a given time point by
348 measuring the differences in the variants produced for different scenes in different rounds. Then we
349 examined changes in input variability over time across conditions. We found that input variability
350 significantly decreased over rounds (Model 13: $\beta=-0.1$, $SE=0.01$, $t=-8$, $p<0.0001$), with a stronger
351 decrease in the larger groups (Model 13: $\beta=-0.08$, $SE=0.2$, $t=-4.42$, $p=0.0001$). Importantly, this analysis
352 also confirmed that larger groups were indeed associated with greater input variability overall (Model 13:
353 $\beta=1.45$, $SE=0.09$, $t=15.99$, $p<0.0001$) – a critical assumption in the literature (8,14,16,39) and a premise
354 for our hypothesis. We also quantified the degree of shared history between participants. Then, we
355 examined the role of input variability and shared history in promoting changes in linguistic structure by
356 using these measures to predict differences in structure scores over consecutive rounds. We found that
357 more input variability at round n induced a greater increase in structure at the following round (Model 14:
358 $\beta=0.015$, $SE=0.003$, $t=4.8$, $p<0.0001$). Similarly, less shared history at round n induced a greater increase
359 in structure at the following round (Model 15: $\beta=-0.017$, $SE=0.004$, $t=-4.18$, $p=0.0004$). When both
360 predictors were combined in a single model, only input variability was significantly associated with
361 structure differences (Model 16: $\beta=0.011$, $SE=0.004$, $t=2.76$, $p=0.012$), while the effect of shared history
362 did not reach significance (Model 16: $\beta=-0.008$, $SE=0.005$, $t=-1.42$, $p=0.17$) – suggesting that input
363 variability was the main driver for the increase in structure scores.

¹ Complex descriptions in the artificial languages could be interpreted as single words with different affixes, or alternatively as different words combined to a sentence (e.g., with a noun describing shape and a verb describing motion). Therefore, in the current paradigm, there is no meaningful distinction between syntactic and morphological compositionality.

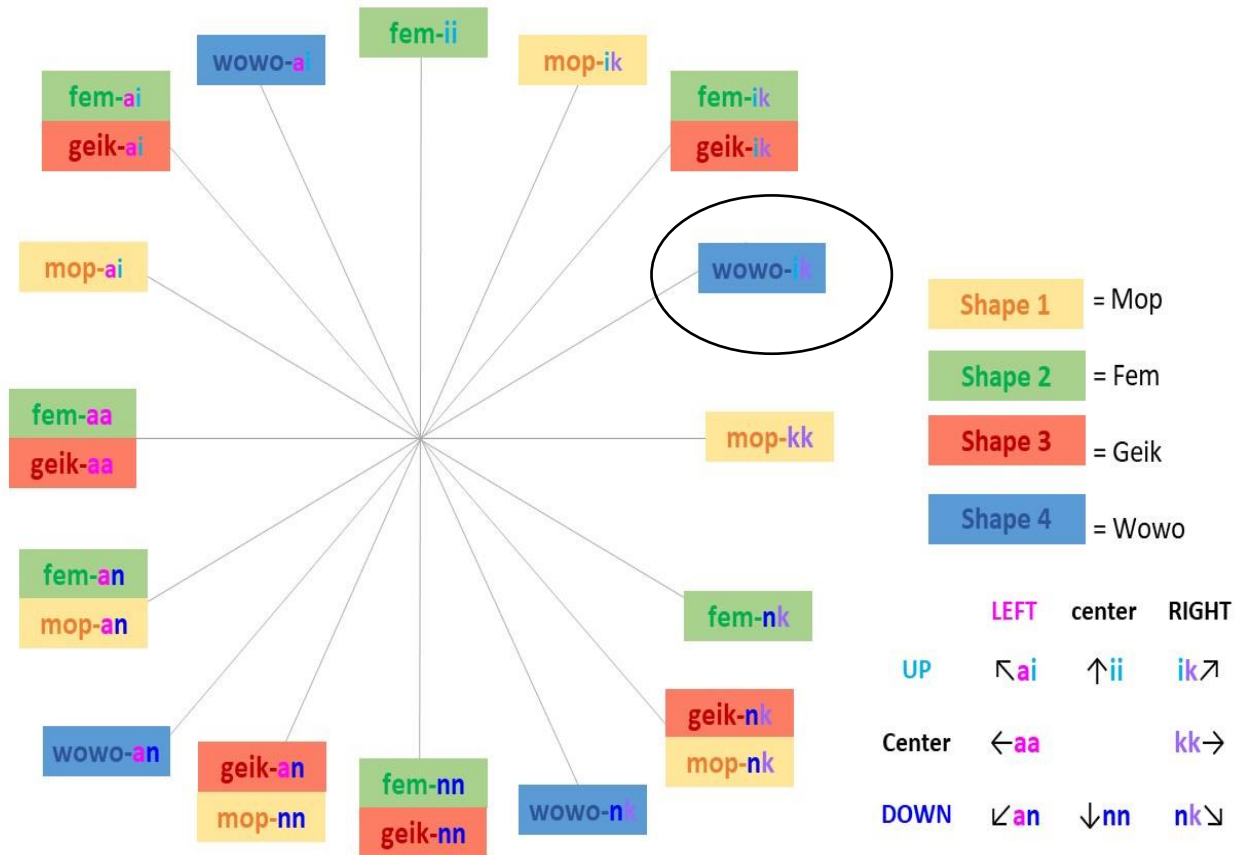


Figure 3. An example of the final language produced by a participant in a large group, along with a “dictionary” for interpreting it on the left. Box colors represent the four shapes, and the grey axes indicate the direction in which the shape moved. Font colors represent different meaningful part-labels, as segmented by the authors for illustration purposes only. For example, the label in the black circle (“wowo-ik”) described a scene in which shape 4 moved in a 30° angle. It is comprised of several parts: “wowo” (indicating the shape) and “ik” (indicating the direction, comprised of two meaningful parts: “i” for “up” and “k” for “right”).

364

Discussion

365 We used a group communication paradigm to test the effect of community size on linguistic structure. We
 366 argued that larger groups were under stronger pressure to develop shared languages to overcome their
 367 greater communicative challenge, and therefore created more systematic languages. We found that while
 368 all larger groups consistently showed similar trends of increasing structure over time, some small groups
 369 never developed systematic grammars and relied on holistic, unstructured labels to describe the scenes.
 370 Importantly, linguistic structure increased faster in the larger groups, so that by the end of the experiment,
 371 their final languages were significantly more systematic than those of small groups. Our results further
 372 showed that the increase in structure was driven by the greater input variability in the larger groups.
 373 Remarkably, the languages developed in larger groups were eventually as globally shared across members,
 374 even though members of larger groups had fewer opportunities to interact with each other, and had more
 375 people they needed to converge with compared to members of small groups. Finally, the languages of
 376 small groups changed less over time, though larger groups reached an equal level of stability by the end
 377 of the experiment. Together, these results suggest that group size can affect the live formation of new
 378 languages.

379 The groups in our experiment were smaller than real-world communities. The results, however, should
380 scale to real-world populations since the meaning space and speakers' life span scale up proportionally.
381 Concordantly, our results are consistent with findings from real developing sign languages, which show
382 that given the same amount of time, a larger community of signers developed a more uniform and more
383 systematic language compared to a small community of signers (14). It also resonates with
384 psycholinguistic findings that show how input variability can affect generalization (22): participants
385 typically don't generalize over variants when they are able to memorize all of them individually, but do
386 generalize when there are too many variants to remember. Similarly, greater input variability in larger
387 groups promoted generalizations of the linguistic stimuli in our experiment, consistent with language
388 change theories that argue for more systematicity in big communities of speakers for the same reasons
389 (8,15-17).

390 The proposed mechanisms assumes a close relationship between our linguistic measures, and is based
391 on the hypothesis that linguistic structure can facilitate convergence and comprehension. We assumed that
392 larger groups compensated for their greater communicative challenge by developing more systematic
393 languages, which enabled them to reach similar levels of convergence and accuracy by the end of the
394 experiment. Therefore, one may wonder whether more structure indeed facilitated convergence and
395 communicative success in our experiment. To this end, we examined the relation between our measures
396 of communicative success, convergence and linguistic structure after controlling for the effect of round
397 (see Appendix C). One model predicted convergence as a function of time and linguistic structure. The
398 model included ROUND NUMBER (centered), STRUCTURE SCORE (centered), and the interaction between
399 them as fixed effects. Another model predicted communicative success as a function of time, convergence,
400 and linguistic structure scores, with fixed effects for ROUND NUMBER (centered), STRUCTURE SCORE
401 (centered), MEAN CONVERGENCE (centered), and the interaction terms STRUCTURE X ROUND and
402 CONVERGENCE X ROUND. Both models included by-group random intercepts and by-group random slopes
403 for all fixed effects. Indeed, we found that more linguistic structure predicted better convergence across
404 different rounds (Model 17: $\beta=0.018$, $SE=0.008$, $t=2.32$, $p=0.027$). Additionally, communicative success
405 was predicted by structure (Model 18: $\beta=0.436$, $SE=0.06$, $t=7.48$, $p<0.0001$) and convergence (Model 18:
406 $\beta=0.189$, $SE=0.06$, $t=2.95$, $p=0.008$), so that better group alignment and more systematic structure
407 predicted higher accuracy scores across rounds. Moreover, the relationship between structure and
408 accuracy became stronger over rounds (Model 18: $\beta=0.051$, $SE=0.008$, $t=6.38$, $p<0.0001$). These
409 additional analyses provide important empirical evidence in support of the underlying mechanisms we
410 proposed, and shed light on the nature of the group size effects reported in this paper.

411 Another important aspect of our results concerns the effect of group size on variance in behavior. We
412 found significantly more variance in the behaviors of small groups across all measures: some groups
413 reached high levels of communicative success, convergence, stability, and linguistic structure, while
414 others did not show much improvement in these measures over time. By contrast, larger groups all showed
415 similar levels of communicative success, stability, convergence, and linguistic structure by the end of the
416 experiment. These results support the idea that small groups are more vulnerable to drift (18,35): random
417 changes are more likely to occur in smaller populations, while larger populations are more resilient to
418 such random events and often show more consistent behaviors. This result may be underpinned by basic
419 probability statistics: small samples are typically less reliable and vary more from each other, while larger
420 samples show more normally distributed patterns and are more representative of general trends in the
421 population ("the law of large numbers" (53)).

422 Our findings support the proposal that community size can drive the cross-linguistic and historical
423 findings that larger societies have more simplified grammars (5,8,14-17), and suggest that differences in
424 community size can help explain and predict patterns and trajectories in language formation and change.
425 Our results show that the mere presence of more people to interact with introduces a stronger pressure for
426 systemization and for creating more linguistic structure, suggesting that an increase in community size

427 can cause languages to lose complex holistic constructions in favor of more transparent and simplified
428 grammars. As such, our results are in line with the idea that increasing community size could have been
429 one of the drivers for the evolution of natural language (19).

430 Our findings also stress the role of the social environment in shaping the grammatical structure of
431 languages, and highlight the importance of examining other relevant social properties alongside
432 community size. Particularly, network structure and connectivity are typically confounded with
433 community size, and have been argued to play an important role in explaining cross-cultural differences
434 in linguistic complexity. Specifically, theories of language change suggest that differences in network
435 density may be the true underlying mechanism behind language simplification (15-17). This idea is
436 supported by computational work showing that networks' structural properties, such as their degree of
437 clustering and hierarchy, can influence linguistic complexity and modulate the effect of population size
438 (21; but see 35). Future work should examine the individual role and mutual influence of these factors to
439 provide a full understanding of how the social environment shapes language evolution.

440

References

- 441 1. Lewis, M.P., Simons, G.F., and Fennig, C.D. (2017). *Ethnologue: Languages of the world* (SIL international Dallas, TX).
- 442 2. Dryer, M.S., and Haspelmath, M. eds. (2017). *WALS Online* (Leipzig: Max Planck Institute for Evolutionary
443 Anthropology) Available at: <http://wals.info/>.
- 444 3. Evans, N., and Levinson, S.C. (2009). The myth of language universals: Language diversity and its importance for cognitive
445 science. *Behav. Brain Sci.* 32, 429–448.
- 446 4. Maffi, L. (2005). Linguistic, Cultural, and Biological Diversity. *Annu. Rev. Anthropol.* 34, 599–617.
- 447 5. Lupyan, G., and Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE* 5, e8559.
- 448 6. Bentz, C., and Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Lang.*
449 *Dyn. Change* 3, 1–27.
- 450 7. Lupyan, G., and Dale, R. (2016). Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity.
451 *Trends Cogn. Sci.* 20, 649–660.
- 452 8. Nettle, D. (2012). Social scale and structural complexity in human languages. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367,
453 1829–1836.
- 454 9. Everett, C., Blasi, D.E., and Roberts, S.G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological
455 and geographic dots. *Proc. Natl. Acad. Sci.* 112, 1322–1327.
- 456 10. Everett, C. (2013). Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives. *PLoS ONE* 8,
457 e65275.
- 458 11. Everett, C., Blasi, D.E., and Roberts, S.G. (2016). Language evolution and climate: the case of desiccation and tone. *J.*
459 *Lang. Evol.* 1, 33–46.
- 460 12. Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., and
461 Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Lang. Learn.* 59, 1–26.
- 462 13. Fusaroli, R., and Tylén, K. (2012). Carving language for social coordination: A dynamical approach. *Interact. Stud.* 13,
463 103–124.
- 464 14. Meir, I., Israel, A., Sandler, W., Padden, C.A., and Aronoff, M. (2012). The influence of community on language structure:
465 evidence from two young sign languages. *Linguist. Var.* 12, 247–291.
- 466 15. Trudgill, P. (2002). Linguistic and Social Typology. In *The Handbook of Language Variation and Change*, J. K. Chambers,
467 P. Trudgill, and N. Schilling-Estes, eds. (Oxford, UK: Blackwell Publishing Ltd), pp. 707–728.
- 468 16. Wray, A., and Grace, G.W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural
469 influences on linguistic form. *Lingua* 117, 543–578.
- 470 17. Milroy, J., and Milroy, L. (1985). Linguistic change, social network and speaker innovation. *J. Linguist.* 21, 339–384.
- 471 18. Nettle, D. (1999). Is the rate of linguistic change constant? *Lingua* 108, 119–136.
- 472 19. Dunbar, R.I.M. (1993). Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* 16, 681–
473 694.
- 474 20. Dale, R., and Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis.
475 *Adv. Complex Syst.* 15, 1150017.
- 476 21. Lou- Magnuson, M., & Onnis, L. (2018). Social Network Limits Language Complexity. *Cogn. Sci.* 42(8), 2790-2817.
- 477 22. Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychol. Sci.* 13, 431–436.
- 478 23. Lev-Ari, S. (2018). The influence of social network size on speech perception. *Q. J. Exp. Psychol.*, 1747021817739865.
- 479 24. Lev-Ari, S. (2016). How the Size of Our Social Network Influences Our Semantic Skills. *Cogn. Sci.* 40, 2050–2064.

- 480 25. Lively, S.E., Logan, J.S., and Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of
481 phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* *94*, 1242–1255.
- 482 26. Bradlow, A.R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* *106*, 707–729.
- 483 27. Perry, L.K., Samuelson, L.K., Malloy, L.M., and Schiffer, R.N. (2010). Learn Locally, Think Globally: Exemplar
484 Variability Supports Higher-Order Generalization and Word Learning. *Psychol. Sci.* *21*, 1894–1902.
- 485 28. Rost, G.C., and McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability
486 in early word learning. *Infancy Off. J. Int. Soc. Infant Stud.* *15*.
- 487 29. Rost, G.C., and McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Dev.*
488 *Sci.* *12*, 339–349.
- 489 30. Scott-Phillips, T.C., and Kirby, S. (2010). Language evolution in the laboratory. *Trends Cogn. Sci.* *14*, 411–417.
- 490 31. Gong, T., Baronchelli, A., Puglisi, A., and Loreto, V. (2012). Exploring the roles of complex networks in linguistic
491 categorization. *Artif. Life* *18*, 107–121.
- 492 32. Wichmann, S., and Holman, E.W. (2009). Population Size and Rates of Language Change. *Hum. Biol.* *81*, 259–274.
- 493 33. Reali, F., Chater, N., and Christiansen, M.H. (2018). Simpler grammar, larger vocabulary: How population size affects
494 language. *Proc. R. Soc. B Biol. Sci.* *285*, 20172586.
- 495 34. Vogt, P. (2009). Modeling interactions between language evolution and demography. *Hum. Biol.* *81*, 237–258.
- 496 35. Spike, M. (2017). Population size, learning, and innovation determine linguistic complexity. *Cog. Sci.* Available at:
497 <https://pdfs.semanticscholar.org/25a0/7b1559b078c07727e0ef1692fb5ae8ebb59e.pdf>.
- 498 36. Gong, T., Shuai, L., and Zhang, M. (2014). Modelling language evolution: Examples and predictions. *Phys. Life Rev.* *11*,
499 280–302.
- 500 37. Galantucci, B., and Garrod, S. (2011). Experimental Semiotics: A Review. *Front. Hum. Neurosci.* *5*, 11.
- 501 38. Roberts, S., and Winters, J. (2012). Social structure and language structure: The new nomothetic approach. *Psychol. Lang.*
502 *Commun.* *16*, 89–112.
- 503 39. Atkinson, M., Kirby, S. and Smith, K. (2015). Speaker input variability does not explain why larger populations have
504 simpler languages. *PLoS one*, *10*(6), e0129463.
- 505 40. Atkinson, M., Smith, K. and Kirby, S. (2018). Adult learning and language simplification. *Cogn. Sci.* *42*(8), 2818–2854.
- 506 41. Atkinson, M., Mills, G.J. and Smith, K. (2018). Social group effects on the emergence of communicative conventions and
507 language complexity. *J. Lang. Evol.* *4*(1), 1–18.
- 508 42. Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interact.*
509 *Stud.* *11*, 138–159.
- 510 43. Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). The Interactive Evolution of Human Communication Systems.
511 *Cogn. Sci.* *34*, 351–386.
- 512 44. Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of
513 linguistic structure. *Cognition* *141*, 87–102.
- 514 45. Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to
515 the origins of structure in human language. *Proc. Natl. Acad. Sci.* *105*, 10681–10686.
- 516 46. Roberts, G., and Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Lang. Cogn.*
517 *4*, 297–318.
- 518 47. Raviv, L., Meyer, A., and Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission.
519 *Cognition* *182*, 151–164.
- 520 48. Spike, M. (2016). Minimal requirements for the cultural evolution of language. PhD thesis, The University of Edinburgh.
- 521 49. Bates, D.M., Maechler, M., Bolker, B., and Walker, S. (2016). lme4: mixed-effects modeling with R.
- 522 50. Halekoh, U., and Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear
523 mixed models—the R package pbkrtest. *J. Stat. Softw.* *59*, 1–30.
- 524 51. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing,
525 Vienna, Austria.
- 526 52. Beckner, C., Pierrehumbert, J.B., and Hay, J. (2017). The emergence of linguistic structure in an online iterated learning
527 task. *J. Lang. Evol.* *2*(2), 160–176.
- 528 53. Blume, J.D., and Royall, R.M. (2003). Illustrating the Law of Large Numbers (and Confidence Intervals). *Am. Stat.* *57*,
529 51–57.