Simple kinship systems are more learnable

Kenny Smith (kenny.smith@ed.ac.uk), Stella Frank, Sara Rolando, Simon Kirby & Jia E. Loy
Centre for Language Evolution, University of Edinburgh, Dugald Stewart Building, Charles Street, Edinburgh, UK

Abstract

Natural languages partition meanings into labelled categories in different ways, but this variation is constrained: languages appear to achieve a near-optimal trade-off between simplicity and informativeness. Across 3 artificial language learning experiments, we verify that objectively simpler kinship systems are easier for human participants to learn, and also show that the errors which occur during learning tend to increase simplicity while reducing informativeness. This latter result suggests that pressures for simplicity and informativeness operate through different mechanisms: learning favours simplicity, but the pressure for informativeness must be enforced elsewhere, e.g. during language use in communicative interaction.

Keywords: language; kinship; complexity

Introduction

Different languages partition meanings into different semantic categories, labelled with words or morphemes. The scope of variation in these partitions is wide, as systems of semantic categories can differ in both the number of labels used and in the strategies used to group meanings into categories. However, this variation is constrained – not all theoretically-possible partitions are found in natural languages, and similar meanings are encountered in unrelated languages.

This pattern of constrained variation has been attested in several domains, such as colour (Berlin & Kay, 1969), number (Greenberg, 1978), and kin classification (Murdock, 1970). Kemp, Xu, and Regier (2018) propose that the constrained variation seen in these systems is a consequence of pressures for efficient communication (see also e.g. Kemp & Regier, 2012). According to this view, category systems are shaped by two competing forces: the need for simplicity (an efficient category system minimises cognitive load), and the need for accurate communication (an efficient category system allows listeners to reliably reconstruct the meanings intended by the speaker). In general, simplicity and informativeness will conflict — the simplest category system (which uses one word for all meanings) is not informative, and the most informative category systems (which divide the world into many fine-grained categories) are maximally complex.

Kemp et al. (2018) suggest that languages exist along an optimal frontier, balancing simplicity and informativeness: natural languages tend to adopt the simplest grammar yielding a given level of informativeness and tend to have maximal informativeness for a given level of complexity. In addition to accounting for fine-grained variation among languages, this same trade-off between simplicity and informativeness has been implicated in the evolution of fundamental structural properties shared by all languages, e.g. combinatorial phonological coding (Oudeyer, 2005) and compositionality (Kirby, Tamariz, Cornish, & Smith, 2015), which also represent optimal trade-offs between simplicity and informativeness.

There is, however, some debate in the literature about the mechanisms which impose these pressures. Kirby et al. (2015) argue that pressures for simplicity are imposed during learning (e.g. in intergenerational transmission), whereas pressures for informativeness apply only during communicative use; as a result, systems which are transmitted from person to person but not employed for communication should become increasingly simple and consequently lose communicative function. In contrast Carstensen, Xu, Smith, and Regier (2015) suggest that pressures for informativeness might operate during learning, such that category systems which are repeatedly learned will become increasingly informative, the opposite prediction to that from Kirby et al. (2015) (see e.g. Fedzechkina, Jaeger, & Newport, 2012 for similar claims).

Here we focus on the case of kinship systems (sets of words used to refer to relationships between family members). Kemp and Regier (2012) show that kinship systems in natural languages exhibit a near-optimal trade-off between informativeness (ability to uniquely specify individuals in a family tree) and simplicity (which they quantify by the length of the underlying grammar). In Experiments 1–2 we test one of the key assumptions in Kemp and Regier's argument, investigating whether simpler, more compressible (artificial) kinship systems are indeed more learnable than less compressible alternatives. We find that objectively simpler kinship systems are indeed easier to learn, and that errors in learning tend to reduce complexity at the cost of decreasing informativeness. In Experiment 3 we explore the informativeness question further, verifying that learners sacrifice expressive power in favour of representational simplicity.

Experiment 1

Participants attempted to learn the language of an imaginary community consisting of 12 members of an extended family. The experiment comprised two phases: character familiarisation followed by the main language task. We varied the complexity of the target kinship system between-subjects.

Methods

Participants Fifty participants (25 per target kinship system) were recruited from the student population at The University of Edinburgh and paid £6 to participate.

Materials: The family tree The community consisted of 12 members of a family tree (see Figure 1A). We trained participants on kinship terms for 16 of the relationships which can be depicted using individuals drawn from this family: maternal grandmother, maternal grandfather, paternal grandmother, paternal grandfather, mother, father, maternal aunt, maternal uncle, paternal aunt, paternal uncle, sister, brother, daughter, son, granddaughter, grandson.

Materials: Languages We generated the initial language for each participant by randomly combining 2–4 CV syllables to produce 12 non-words (e.g. *walo*, *pugowo*, *kohuhake*). These 12 labels were used to express the 16 relationships listed above: eight labels referred to unique relationships; the remaining four were homonyms that referred to two possible relationships. Initial languages were either *simple* or *complex*, depending on the placement of homonyms. Simple languages used homonyms for the pairs <maternal grandfather, paternal grandfather>, <brother, sister>, <maternal uncle, maternal aunt>, and <paternal uncle, paternal aunt>; complex languages used homonyms for the pairs <maternal grandmother, paternal aunt>, <maternal grandfather, paternal uncle>, <paternal grandmother, maternal aunt>, and <paternal grandfather, grandson>. Thus, simple languages assigned identical labels to similar meanings which could be grouped under the same category (e.g., "mother's sibling"), while complex languages assigned identical labels to very different meanings. The two kinship systems differ on the objective measure of complexity used by Kemp and Regier (2012) (i.e. it takes a larger grammar to capture the complex kinship system) and also in the measure of complexity we develop below.

Procedure: Character familiarisation Participants were familiarised with the members of the community and their relationships with each other through a series of introduction and test items. Introduction items presented images of family members in groups of three and stated their relationship, e.g. *This is Mimi and Gonn. Mimi and Gonn have a child: Lulu*. Each introduction item was followed by a test item which tested participants on the relationship they had just been exposed to, for example presenting one character and asking participants to select another character who was parent or child of that character (e.g. *Who is Lulu's parent?*). No English kinship terms were used during the familiarisation phase except the primitives "parent" and "child". Familiarisation consisted of 8 introductory–test pairs and a further 10 test items. Participants were not allowed to proceed to the language task until all test items had been answered correctly; test items which were answered incorrectly were re-presented immediately.

Procedure: Language task Participants were simultaneously trained and tested on the kinship system of the community. Participants were told their goal was to learn how members of the community greeted one another. Participants saw family members greet other family members with the greeting *luha* and then a kinship term (i.e. the greetings were equivalent to e.g. *hello father!*, *hello auntie!*). This phase presented two types of trials: production (Figure 1B) and comprehension (Figure 1C). After each trial, participants received feedback on the accuracy of their response as well as the correct answer if their response was incorrect (except in the final block of testing–see below).

The language task comprised five blocks of testing with 32 trials per block (16 production trials, 16 comprehension trials, alternating). Each relationship (mother, father, grandfather etc) was represented twice per block, once in a production trial and once in a comprehension trial, order randomised. To ensure that participants could remember the family tree, after the first and second blocks of testing participants saw 5 test items similar to those in the familiarisation phase. As before, these were presented repeatedly until participants selected the correct response. On the final block of testing, participants received no feedback on the accuracy of their responses, providing us with a final language from each participant for analysis.

Kinship System Inference

We infer each participant's underlying kinship system from their productions in order to assess their complexity. We use the model from Mollica and Piantadosi (under revision) to sample possible kinship rules for each label used by the participant, then measure the (summed) complexity of those rules and the compressibility of the set of rules.

Rules are set functions, taking the speaker as input and returning a set of individuals who are possible referents for the kinship term used in the greeting. The functions are constructed using a small set of base functions, namely: PARENT(X), CHILD(X), MALE(X), FEMALE(X), UNION(X,Y). The term 'mother' could thus be represented as FEMALE(PARENT(X)); however, it is important to note that there are many other possible rules that, if Nene (see Figure 1 for character names) is the speaker, include Kiki in the set of possible referents. For example: PARENT(X); CHILD(PARENT(PARENT(X))); or UNION(PARENT(X), PARENT(PARENT(X))).

The model defines the Bayesian posterior probability of each possible rule as a combination of the rule's simplicity (in the prior) and its precision (in the likelihood). The simplicity prior $P(r)$ is the product of the probability of each of the rule's components, and thus relates to the number of base functions used to define the rule. The size principle likelihood $P(d|r)$ evaluates the number of possible referents given a data point (speaker, referent, word): the larger the set of possible referents, the lower the likelihood (Tenenbaum & Griffiths, 2001). In the above 'mother' example, the broad PARENT(X) rule, applied to Nene as the speaker and Kiki as the intended referent, would have a high prior probability due to its simplicity,

Figure 1: A. Individuals in the family tree. Members within each branch shared physical features to aid the recognition of relatedness. B. Example production trial: the participant sees the family members involved and must select the appropriate kinship term (in the example depicted, in English the appropriate greeting would be *"Hello grandmother"*). C. Example comprehension trial: the participant sees the speaker and their choice of kinship term and has to select the appropriate addressee from the set of all family members (e.g. in the trial depicted, if *nulenage* means the equivalent of *mother* in English, the participant should select Kiki, who is Nene's mother).

but a lower likelihood as it picks out two possible referents, in contrast to the FEMALE(PARENT(X)) rule.

We infer participant lexicons/kinship systems as follows. For each participant, we use the final set of productions as data, duplicated (i.e., doubled) to increase the pressure on likelihood. We then use the Metropolis-Hasting sampler described in Mollica and Piantadosi (under revision)[1] to generate a set of high-probability candidate rules for each term, given the participant's productions. In order to keep inference time feasible and to counteract the paucity of data, the sampler is initialised with the rules which generated the participant's input language, which allows us to run the sampler for a relatively small number of steps (4000 steps across 40 chains) and still generate acceptable results. Due to the initialisation procedure, the sampler will preferentially explore rules that are similar to the rules used to generate the training data, albeit with high likelihood under the participant's productions, which often do not fit the training data very closely.

Lexicons are then created by drawing a rule for each term from the set of sampled rules, in proportion to the posterior probability of that rule. We create 100 lexicons for each participant, to evaluate the distribution over probable lexicons for that participant, rather than choosing a single representative lexicon. Each lexicon is evaluated along two criteria: the product of the prior probability of each of the rules in the lexicon, and its compressibility (using Lempel-Ziv compression on the inferred lexical meanings[2]). Both the prior and the lexicon compression measure evaluate the complexity of the kinship system. The prior probability of the inferred lexicon (i.e. the product of the prior probabilities of the individual rules) evaluates the simplicity of each of the individual rules, while lexicon compression measures reuse across rules.

Under this measure the simple and complex input languages differ in their complexity, both on the prior probability of their lexical rules and the compressibility of the entire lexicon: we inferred 100 lexicons for each input language, the lexicons for the simple input language have mean log prior probability of -128 and compressed size of 253 bits, the lexicons for the complex input language have log prior probability of -214 and compressed size of 329 bits.

Results

We analysed participants' accuracy (the proportion of trials where participants clicked on the correct response) across all 5 blocks of the experiment, the communicative functionality of the languages they produced in the final testing block (as measured by the number of labels and communicative cost), and the complexity of the kinship system produced in the final testing block (as measured by the prior probability and compressibility of their inferred lexicon). We expected that 1) participants learning the simple system would reach higher accuracy in a shorter time than participants learning the complex system; 2) participants trained on simple kinship systems would produce simpler systems on test; 3) participants' final languages would be simpler and less communicatively functional than their input, reflecting a general bias in learning in favour of simplicity at the expense of informativeness.

Accuracy Figure 2A shows participants' accuracy (proportion of trials on which participants clicked on the correct response) over blocks in production and comprehension trials. We used logistic mixed effects models to analyse the binary outcome of participants' response on each trial (correct/incorrect), with input complexity, block, and their interaction as fixed effects;[3] we ran separate models for pro-

[1] https://github.com/MollicaF/LogicalWordLearning.

[2] This involves encoding each rule as a list of integers, where each term in each rule is encoded as a single integer, with each instance of a term receiving the same code. Each integer then gets turned into a bitstring using Fibonacci coding, and the resulting bitstring is compressed using the LZ2algorithm. This metric is provided in the code accompanying Mollica and Piantadosi (under revision).

[3] Input complexity was deviation coded; we coded block such that the model intercept reflected performance at the final test block. Models included random intercepts for subjects and by-subject random slopes for block.

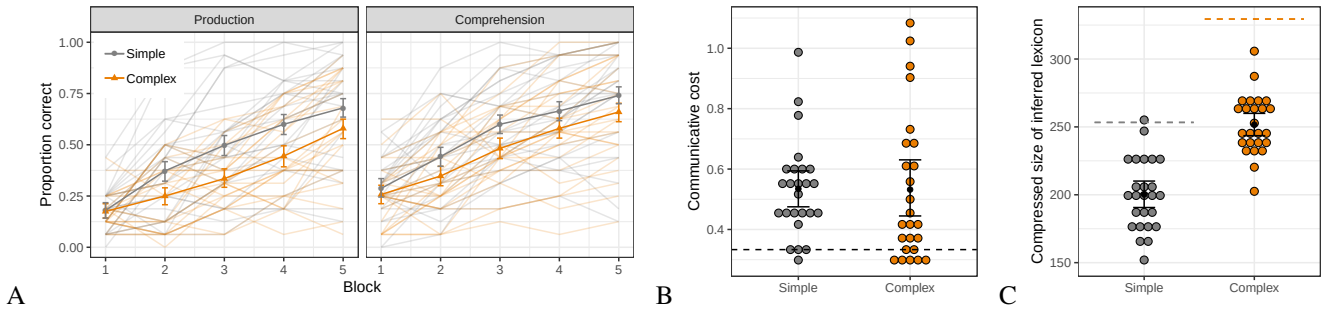

Figure 2: Experiment 1 results. A. Performance over time in production and comprehension. Solid lines give means and bootstrapped 95% CIs; individuals are also shown as fainter lines. B. Communicative cost of participants' final languages. Dashed lines show the input, large points indicate individual participants, small point and error bars give means and 95% CIs. C. Compressed size of participants' final languages. Plotting conventions as in B.

duction and comprehension trials due to the different nature of the two tasks. Participants' performance improved over time: the models for accuracy showed a significant effect of block in both production ($b = 0.58$, $SE = 0.05$, $p < .001$) and comprehension trials ($b = 0.57$, $SE = 0.05$, $p < .001$). The model fitted to production trials suggests a highly marginal interaction of complexity and block ($b = -0.17$, $SE = 0.10$, $p = .093$), which is consistent with slower learning of the complex system; this results in a significant effect of complexity at block 5, with the participants trained on complex input having significantly lower production accuracy on block 5 ($b = -0.96$, $SE = 0.42$, $p = .023$). These effects are n.s. in the comprehension model ($p > .123$), suggesting clearer effects of complexity on production than comprehension.

Communicative function of final languages We assessed communicative functionality of the final languages produced by our participants (i.e. on production trials in block 5) according to two measures: the number of distinct labels produced (fewer labels will usually lead to a drop in communicative function as labels become increasingly ambiguous) and the communicative cost of the labels produced (plotted in Figure 2B: communicative cost for a label L is $-log_2(1/|L|)$ where $|L|$ is the number of relationships L applies to; the communicative cost of a lexicon is the average cost of its labels[4]).

We used regression to analyse the number of labels and communicative cost (Poisson regression for the former), with input complexity (deviation coded) as a fixed effect. There was no difference between input conditions for number of labels (mean for simple lexicons, M_{simple}=10.28 labels; $M_{complex}$=10.2 labels; $b = 0.01$, $SE = 0.09$, $p = .930$) or communicative cost ($b = 0.00$, $SE = 0.06$, $p = .967$). Participants on average produced fewer labels than in their input (log number of labels in the model intercept is significantly lower than log(12), $b = 2.33$, $SE = 0.04$, $p < .001$), yielding languages with higher communicative cost (model intercept is higher than the communicative cost of the input languages, 1/3: $b = 0.53$, $SE = 0.03$, $p < .001$).

Complexity of final languages We assessed the complexity of the final languages produced by our participants according to the two measures introduced above: the summed log prior probabilities of the lexical items (higher prior probability indicates lower complexity) and the size of the compressed set of rules (smaller size indicates a more compressible, simpler rule system; see Figure 2C).

We used regression to analyse lexicon prior probability and compressed size, with input complexity (deviation coded) as a fixed effect. There was a significant difference between input conditions for both lexicon prior probability (mean log prior for simple condition $= -98$, mean for complex condition $= -130$; $b = -32.73$, $SE = 4.28$, $p < .001$) and compressed size (mean size for simple condition $= 200$ bits, mean for complex condition $= 252$ bits; $b = 51.43$, $SE = 6.79$, $p < .001$), with participants trained on the complex system producing less compressible lexicons with more complex rules. As can be seen in Figure 2C, participants in both conditions produce simpler systems than they were trained on (the difference between input and output complexity is significant at $p < .001$ on both measures even for the simple input condition), but this difference is clearly far larger for participants in the complex input condition.

Experiment 1 discussion

These results are consistent with the theory that learning favours simpler languages, and that errors made during learning tend to decrease complexity while also reducing communicative utility, i.e. increasing communicative cost. However, our task was quite challenging and produced substantial inter-individual variation (as seen in e.g. the accuracy over time shown in Figure 2A), and some effects are marginal (most notably the difference in learning rates across conditions); we therefore attempted to replicate these results.

[4]Kemp and Regier (2012) include a weighting based on need probability in the measure of communicative cost, i.e. the probability of being required to communicate about different relationships; in our experiment each relationship is labelled equally frequently, making the need probabilities (at least in the context of the experiment) uniform, allowing this term to be dropped.

Experiment 2

Experiment 2 is a replication of Experiment 1; to facilitate rapid collection we ran the experiment online.

Methods

Participants Eighty participants (40 per target kinship system) were recruited using Amazon Mechanical Turk and paid $6 to participate.

Materials, procedure, analysis Identical to Experiment 1.

Results

Accuracy Figure 3A shows participants' accuracy over blocks in production and comprehension trials. As in the lab experiment, participants' performance improved over time: there was a significant effect of block in both production ($b = 0.32$, $SE = 0.03$, $p < .001$) and comprehension trials ($b = 0.37$, $SE = 0.03$, $p < .001$). The model fitted to production trials shows a significant interaction between complexity and block ($b = -0.14$, $SE = 0.07$, $p = .035$), indicating slower learning of the complex system (recall this effect was highly marginal in Experiment 1); this difference in learning rates results in a significant effect of complexity at block 5 ($b = -0.71$, $SE = 0.27$, $p = .009$). As in Experiment 1 the equivalent effects are n.s. in comprehension trials ($p > .128$).

Communicative function of final languages As in Experiment 1, there was no difference between conditions for number of labels ($M_{simple} = 9.62$; $M_{complex} = 8.85$; $b = 0.08$, $SE = 0.07$, $p = .254$), but there was a marginal difference in communicative cost, with participants trained on complex input perhaps producing languages with higher cost ($b = -0.09$, $SE = 0.05$, $p = .077$; see Figure 3B). As in the lab, online participants produced languages with fewer labels ($b = 2.22$, $SE = 0.04$, $p < .001$) and higher communicative cost ($b = 0.65$, $SE = 0.02$, $p < .001$) than their input.

Complexity of final languages As in Experiment 1 there was a significant difference between input conditions for both lexicon prior probability ($M_{simple} = -111$; $M_{complex} = -121$; $b = -10.79$, $SE = 3.47$, $p = .003$) and compressed size ($M_{simple} = 225$; $M_{complex} = 244$; $b = 19.80$, $SE = 6.9$, $p = .002$, see Figure 3C), with participants trained on the complex system again producing less compressible lexicons with more complex rules. As in Experiment 1, participants in both conditions produce simpler systems than they were trained on (the difference between input and produced complexity is significant at $p < .001$ on both measures even for simple input participants), and this difference is again far larger for participants in the complex input condition.

Experiment 2 discussion

The results from Experiment 2 confirm those of Experiment 1: participants trained on simpler kinship systems learn more rapidly and more accurately than participants attempting to learn complex kinship systems, and as in Experiment 1 errors in learning tend to reduce complexity, particularly for complex input systems, and increase communicative cost. These results are therefore consistent with the view that pressures for simplicity and informativeness come from different mechanisms (i.e. learning and use respectively), or at least are not both at play in learning (contra e.g. Fedzechkina et al., 2012; Carstensen et al., 2015). However, our Experiments 1–2 are a rather unfair test of the idea that learning might be biased in favour of informativeness: participants were trained on a 12-label kinship system, and on test only had 12 labels to select among, meaning they could not introduce new distinctions and straightforwardly reduce communicative cost. Reducing communicative cost *is* possible by redistributing homonyms (e.g. by overloading one homonymous term to refer to 3 individuals, creating a new unambiguous term), but this offers only a modest decrease in communicative cost. As a result our finding that errors in learning reliably increase communicative cost might just reflect a ceiling effect. Furthermore, the 12-label systems we used seem to be at or beyond the capacity for most participants to learn accurately in the time available; in particular, participants typically failed to produce all 12 available labels, therefore inevitably increasing the communicative cost of the system; while this is part of the effect we are interested in, representing a bias for simplicity in learning at the expense of communicative function, it would be worthwhile to verify that a similar bias can be seen in a kinship system featuring a more manageable number of labels. In Experiment 3 we therefore train participants on an input language which uses fewer labels and which offers more scope for learners to restructure their input so as to reduce communicative cost and improve communicative function.

Experiment 3

We trained participants on an input language which uses only 8 distinct labels, which could be glossed in English as *sister*, *brother*, *child* (i.e. picking out the pair of relationships <daughter, son>), *mother and her siblings* (<mother, maternal aunt, maternal uncle>), *father and his siblings* (<father, paternal aunt, paternal uncle>), *maternal grandparent* (<maternal grandmother, maternal grandfather>), *paternal grandparent* (<paternal grandmother, paternal grandfather>), and *grandchild* (<granddaughter, grandson>).

Methods

Participants 41 participants were recruited using Amazon Mechanical Turk and paid $6 to participate.

Materials, procedure, analysis Identical to Experiment 2, with the exception of the input language; note that participants had 12 labels available on production test trials, the 8 labels featured in their training language and 4 'spare' labels.

Results

Accuracy Figure 3A shows participants' accuracy over blocks. As expected, accuracy improves over time in both

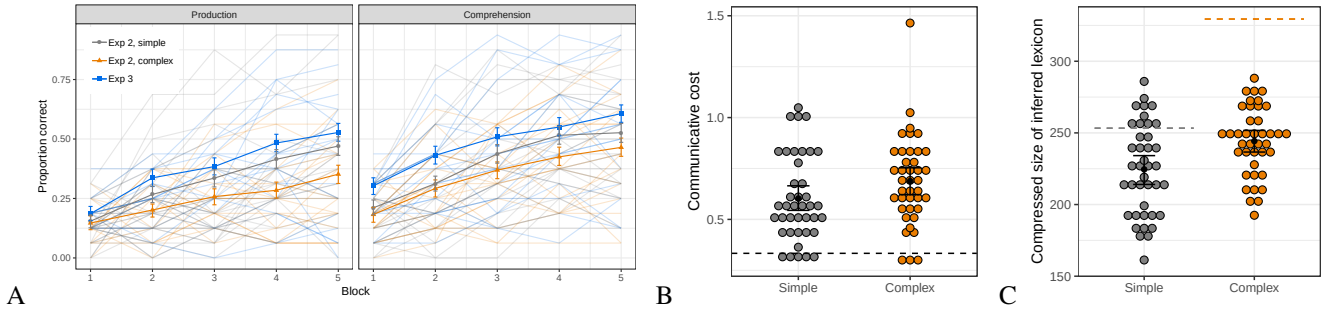

Figure 3: Experiment 2 results. A. Performance over time (Experiment 3 data also plotted; 20 randomly-selected participants per condition shown as individual lines to reduce clutter). B. Number of labels. C. Communicative cost.

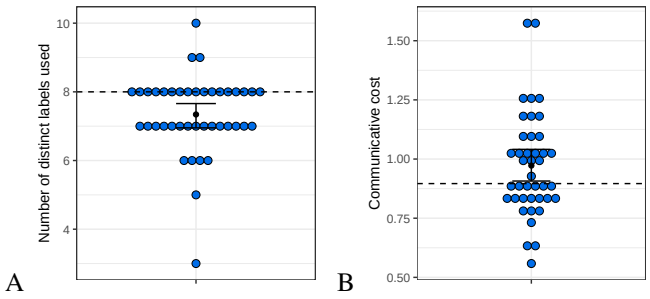

Figure 4: Experiment 3 results. A. Number of labels; B. communicative cost. Accuracy results shown in Figure 3A.

production ($b = 0.35$, $SE = 0.04$, $p < .001$) and comprehension ($b = 0.33$, $SE = 0.04$, $p < .001$); analysing data from Exps 2–3 combined, with condition treatment-coded, indicates that final accuracy in block 5 in Exp 3 is not significantly higher than in simple condition of Exp 2 on either production or comprehension ($p > .44$), but *is* higher than the complex condition of Exp 2 (production: $b = -0.87$, $SE = 0.24$, $p < .001$; comprehension: $b = -0.67$, $SE = 0.23$, $p = .004$).

Communicative function of final languages Figures 4A–B shows the number of distinct labels produced and the communicative cost of the resulting lexicons. While participants do not produce significantly fewer distinct labels than in their input ($b = 1.99$, $SE = 0.06$, $p = .136$), the final languages did however have higher communicative cost than the input language ($b = 0.97$, $SE = 0.03$, $p = .026$).

Experiment 3 discussion

Experiment 3 removes the potential ceiling effect in Experiment 2; it would be possible to redesign the input language to reduce communicative cost, e.g. by using the 'spare' labels to introduce additional distinctions, or by redistributing homonymy more evenly across the 8 labels used in the input. Most of our participants do not do this; instead, as in Experiments 1–2, on average they produce final languages which have even higher communicative cost than their input, again consistent with the view that pressures for informativeness are *not* at play in learning (contra e.g. Fedzechkina et al., 2012; Carstensen et al., 2015). It is also worth noting that the language in Experiment 3 is learnt significantly better than the low-communicative-cost complex language from Experiment 2; this would be surprising if learners were biased against languages with high communicative cost.

General discussion

Across 3 experiments we find that simpler kinship systems are easier to learn for human participants, validating a crucial assumption in Kemp & Regier's (2012) analysis of natural language kinship systems, and matching similar results in other domains indicating that biases in learning favour simplicity (e.g., Feldman, 2000; van de Pol, Steinert-Threlkeld, & Szymanik, 2019; see Feldman, 2016 for review).

Our results are inconsistent with claims that biases in learning instead favour informativeness or communicative efficiency e.g., Carstensen et al. (2015); Fedzechkina et al. (2012). How can we reconcile this difference? Carr, Smith, Culbertson, and Kirby (in press) note that while simplicity and informativeness are often opposed (e.g. having few labels is simple but not informative), there are cases where the biases coincide: in particular, they show that simplicity and informativeness can both favour contiguous categories, where closely-related meanings fall into the same category. Contiguous categories are simple (they can be represented compactly) but also more informative than non-contiguous categories, in that they direct the receiver of the category label to the right region of the semantic space, even if they fail to pick out exactly the right meaning. Carr et al. (in press) show that this can account for the puzzling results from Carstensen et al. (2015), where the apparent increase in informativeness occuring over generations of learning is likely to be driven by a simplicity-based preference for category contiguity. It remains to be seen whether similar alternative explanations exist for other findings suggesting communicative biases in learning, e.g. those in Fedzechkina et al. (2012). It is also worth noting that the measure of communicative cost we use here merely depends on the probability of selecting the correct individual, rather than rewarding near misses (e.g. providing partial payoff for selecting an individual who is sim-

ilar to the target in terms of generation or sex); it may be that some of the errors made by our participants reduce communicative cost if measured in this way, as a side-effect of increasing simplicity.

Another possibility is that the preference for simplicity we see in our experiments merely reflects poor learning, and that we would see a preference for decreased communicative costs if participants were given more time to learn the target systems more accurately. This strikes us as unlikely. First, while many of our participants do indeed have quite low accuracy even by the end of the experiment, it is worth noting that inaccurate learning is a necessary feature of our design, rather than a flaw: some errors are necessary in order to reveal biases in learning, and if we trained participants to perfect accuracy on the target systems we would not be able to measure their deviations from those target systems. Second, in order for different biases to be seen later in learning there would need to be a discontinuity in the trajectory followed by our learners. Our results suggest that participants (at least initially) approach the target language complexity from below, i.e. via intermediate systems that are lower in complexity; while it is logically possible that later in learning participants would reliably overshoot the target language complexity and then approach from above (i.e. from higher complexity, lower cost systems) before finally settling on the target system, we see no reason to expect this. Third, as mentioned above there is already a wealth of independent evidence suggesting that learning in multiple domains favours simplicity, a pattern of results which our findings are consistent with and expected under. Finally, even if this kind of trajectory was witnessed, we would expect the effects of late-in-learning biases to be quite subtle (because participants are close to the target system), making them harder to spot experimentally but also limiting their impact on the structure of linguistic systems relative to other pressures, such as simplicity biases earlier in learning, or pressures arising during communication.

Conclusion

We provide evidence suggesting that simpler kinship systems are easier to learn for human participants, and that biases in learning favour simple systems even at the expense of communicative function; participants tend to produce simpler kinship systems than they were trained on, and in all experiments the kinship systems they produced would be less communicatively effective (i.e. had higher communicative cost) than their input. This is consistent with accounts where the pressures for simplicity and informativeness which shape natural languages (in their kinship systems and beyond) arise through distinct processes, i.e. learning and use respectively, rather than both being the product of biases in learning.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 681942). Many thanks to Frank Mollica for his assistance in adapting the code from Mollica and Piantadosi (under revision) to our experimental data. Thanks to Charles Kemp and two anonymous reviewers for helpful suggestions.

References

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. University of California Press.

Carr, J., Smith, K., Culbertson, J., & Kirby, S. (in press). Simplicity and informativeness in semantic category systems. *Cognition*.

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012, October). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897–17902.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Feldman, J. (2016). The simplicity principle in perception and cognition: The simplicity principle. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, 330–340.

Greenberg, J. H. (1978). Generalizations about Numeral Systems. In J. H. Greenberg, C. H. Ferguson, & E. A. Moravcsik (Eds.), *Universals of Human Language* (Vol. 3: Word Structure, pp. 249–295). Stanford, CA: Stanford University Press.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018, January). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Mollica, F. M., & Piantadosi, S. T. (under revision). Logical word learning: The case of kinship.

Murdock, G. P. (1970, April). Kin Term Patterns and Their Distribution. *Ethnology*, *9*, 165.

Oudeyer, P.-Y. (2005). The self-organization of combinatoriality and phonotactics in vocalization systems. *Connection Science*, *17*, 325–341.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

van de Pol, I., Steinert-Threlkeld, S., & Szymanik, J. (2019). Complexity and learnability in the explanation of semantic universals of quantifiers. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.