Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions

Word meanings evolve to selectively preserve distinctions on salient dimensions

Catriona Silvey*, Simon Kirby, Kenny Smith

School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, United Kingdom

* C.A.Silvey@sms.ed.ac.uk

Running head: Cultural evolution of word meanings

Keywords: attentional learning; cultural transmission; iterated learning; language evolution; word meaning; language and conceptualization

Abstract

Words refer to objects in the world, but this correspondence is not one-to-one: each word has a range of referents that share features on some dimensions, but differ on others. This property of language is called underspecification. Parts of the lexicon have characteristic patterns of underspecification: e.g., artifact nouns tend to specify shape, but not color, whereas substance nouns specify material but not shape. These regularities in the lexicon enable learners to generalize new words appropriately. How does the lexicon come to have these helpful regularities? We test the hypothesis that systematic backgrounding of some dimensions during learning and use causes language to gradually change, over repeated episodes of transmission, to produce a lexicon with strong patterns of underspecification across these less salient dimensions. This offers a cultural evolutionary mechanism linking individual word learning and generalization to the origin of regularities in the lexicon that help learners generalize words appropriately.

1. Introduction

Language allows us to communicate about the world. This is possible because parts of language (e.g., words) refer to parts of the world (e.g., objects). However, this relationship is rarely one-to-one. For example, the word 'cat' refers to a range of objects that share features on certain dimensions, such as shape, but differ on others, such as color. This abstraction over features is a ubiquitous property of natural language called underspecification (Geeraerts, 2009, p.196).

Different areas of the lexicon have different characteristic patterns of underspecification. For example, words for artifacts tend to specify shape or function, and underspecify color; words for substances tend to specify material, and underspecify shape (Smith & Samuelson, 2006). These regularities in the lexicon enable learners to acquire higher-order generalizations about which dimensions are relevant to the meaning of words learned in particular contexts, for example the shape bias that labels for objects generalize by shape (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

However, this account does not explain how the lexicon comes to have these helpful regularities in the first place. One possibility is that learners have strong constraints on the kind of word meanings they will entertain (Markman, 1994; Waxman & Kosowski, 1990), which map straightforwardly to strong constraints on the kinds of underspecification lexicons can exhibit. Instead, we show that the same processes that enable learners to form higher-order generalizations on the basis of regularities in the

lexicon can also shape the lexicon to exhibit those regularities in the first place, leading it to reflect the systematic salience of particular dimensions in contexts of learning and use. This happens not over the course of an individual's learning, but via the cumulative language change that occurs when a lexicon is transmitted.

The attentional learning account states that 'context cues that co-occur with (and define) specific tasks will come with repeated experience to shift attention to the task-relevant information' (Smith, Colunga, & Yoshida, 2010, p. 1295). Modeling the learning of (part of) the lexicon as this kind of 'specific task', we train and test learners on an artificial language in contexts where one dimension of meaning is systematically made less salient (backgrounded). We manipulate salience by casting word learning and use as a series of discrimination games where one dimension is never helpful. This has a precedent in the 'guessing game' of Steels (2003), along with the well-established results in the concepts and categories literature showing that dimensions that are unhelpful for discrimination are attended to less than helpful dimensions (e.g., Kruschke, 1992; Medin & Schaffer, 1978). In real word learning, this backgrounding effect is more likely the outcome of factors such as domain-specific knowledge (Kelemen & Bloom, 1994; Lin & Murphy, 1997), increased salience of functional features (Booth & Waxman, 2002; Keil, 1994; Kemler Nelson, 1995), attentional cues from speakers (Tomasello, 2000), inference of the speaker's intention (Bloom, 2000; Xu & Tenenbaum, 2007), or other 'non-linguistic evidence of the speaker's locus of attention' (Clark, 1997, p.7).[1] This systematic backgrounding has only a small effect at the individual level. However, over cultural transmission, a lexicon that initially specifies equally across all dimensions changes to reflect the

[1] Some of these factors concern the intrinsic salience of particular object features, rather than (as modeled in this experiment) task-defined salience in situations of learning and use. Intrinsic salience could also have a strong effect in directing underspecification, via the same mechanisms of cultural evolution modeled here.

differing salience of dimensions in learning and use, leading to an emerging system which preferentially underspecifies the backgrounded dimension. This serves as a demonstration of how cultural transmission amplifies the effects of individual learning processes to create an adaptively specified lexicon.

1.1 Modeling the cultural evolution of underspecification: iterated learning

We model the cultural evolution of language using **iterated artificial language learning** (Kirby, Cornish & Smith, 2008; Smith & Wonnacott, 2010). In the diffusion chain instantiation of this paradigm, participants are organized into chains of transmission: an initial language is taught to the first learner in each chain, who subsequently attempts to reproduce that language; this reproduction is then given as learning input to the next participant in the chain, and so on. Using this methodology, researchers have demonstrated the cultural emergence of properties of language including arbitrariness (Caldwell & Smith, 2012; Fay, Garrod, Roberts, & Swoboda, 2010; Theisen-White, Kirby, & Oberlander, 2011), regularity (Reali & Griffiths, 2009; Smith & Wonnacott, 2010), categorization that reflects discontinuities in world structure (Perfors & Navarro, 2011), compositional structure (Kirby et al., 2008, Exp 2; Theisen-White et al., 2011), and underspecification (Kirby et al., 2008, Exp 1).

Our method here is based on Exp 1 from Kirby et al. (2008). The 'meanings' in this study were a series of images that varied in shape (square, circle, triangle), color (black, blue, red) and motion (horizontal, bouncing, spiraling). Each chain was initialized with a language which provided a unique word for each of the 27

meanings: i.e. it specified fully across all dimensions. However, due to the difficulty of accurately learning and reproducing this language given the amount of training provided, participants began to reuse words for referents that differed on certain dimensions. This led, over several generations of transmission, to the emergence of underspecification as a solution to the learning problem: for example, in one chain, every bouncing square came to be labeled ‘tupim’, regardless of color.

However, this underspecification was not consistently directed to any particular dimension. Across the different chains, some languages underspecified color, some shape, and some motion (Cornish, 2011), presumably because, in the learning and testing procedures used in Kirby et al. (2008), no particular dimension was made more or less salient. By contrast, in real word learning and use, some dimensions have higher salience than others (Clark, 1993; Regier, 2005). For particular groups of referents, commonalities across these situations of learning and use will result in certain dimensions being foregrounded and others backgrounded, as per the attentional learning account (Smith, Colunga, & Yoshida, 2010). Our hypothesis is that these systematic differences in dimension salience during individuals’ learning and production can lead, over cultural transmission, to a pattern of underspecification that reflects these differences – a helpful lexicon that aids subsequent learners in making the right kinds of generalizations. In order to test this hypothesis, we ran a modified version of the Kirby et al. (2008) paradigm, where the learning and production procedures are structured to systematically background one meaning dimension: meanings are presented in pairs that share a feature on one consistent dimension, such that attending to this dimension will never help participants discriminate between the two meanings (Fig. 1). The hypothesis is that underspecification will gradually arise on the backgrounded dimension, thus showing

that strong constraints on learners' word meaning hypotheses are not necessary to explain the patterns of underspecification we see in natural language. If, on the other hand, underspecification were to arise indiscriminately on all dimensions (as in Kirby et al., 2008), this would suggest that stronger constraints are needed to explain real-world patterns.

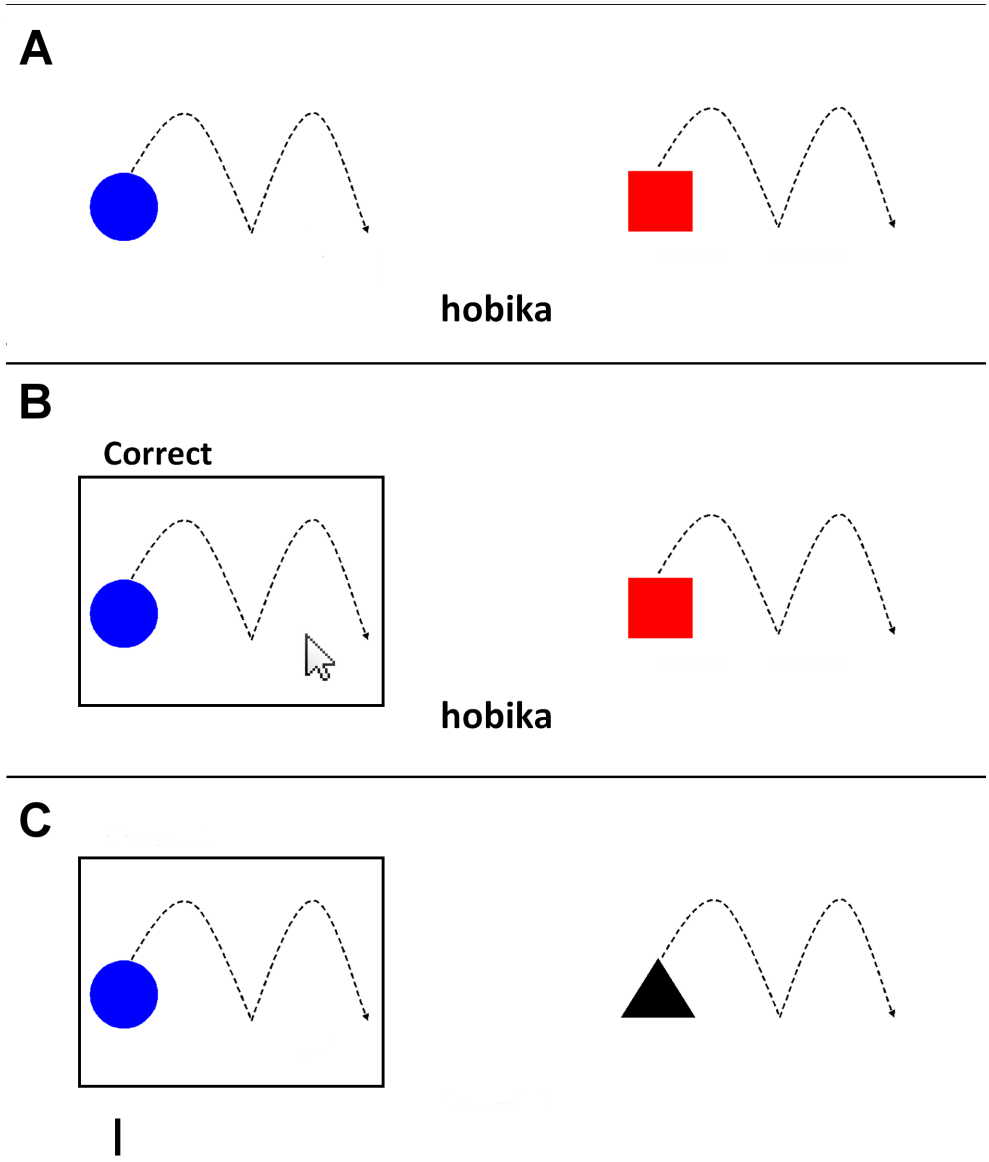


Figure 1. Training and testing procedures in the experiment. A) Each training trial is presented as a discrimination game. The participant is shown a word and two candidate images. The participant clicks the image they think goes with the word. B) The participant is then given feedback, followed by the correct word-image pairing.

The word then disappears and they are required to retype it. C) Test trials are again presented as a discrimination game, but from the opposite perspective. The participant is presented with two images, one of which is selected as the target. They are instructed to type the word that would allow the alien to pick the correct image. In all training and test trials, target and distractor share a feature on one consistent dimension (in this example, the motion dimension).

2. Method

2.1 Participants

40 undergraduate and graduate students at the University of Edinburgh (25 female, median age 20.5) were recruited via mailing lists and organized into 8 diffusion chains. Each chain consisted of an initial participant who was trained on a random language, and 4 successive participants who were trained on the previous participant's test output language, making 5 generations in total: the results of Kirby et al. (2008) suggest that 5 generations would be sufficient for underspecification to arise (in 3 out of their 4 chains the languages had fewer than 5 words by generation 5). Participants in chains 1-6 were unpaid volunteers; participants in chains 7-8 were paid £4.50.[2]

[2] To ensure that the payment of the last two chains of participants did not affect the results, Chain (i.e. which of the 8 chains of 5 learners a participant belonged to) was modeled as a fixed effect in initial analyses to check if this improved the fit of the models. In all cases, the models including Chain as a fixed effect either did not improve overall fit or showed that no particular chain(s) had a significant effect on the results. In the final analyses below, Chain is modeled as a random effect.

2.2 Stimuli: images and input language

Participants were asked to learn and then produce an 'alien language', consisting of lowercase text labels paired with images. The images were the 27 pictures of colored shapes in motion from Kirby et al. (2008). The images varied in three possible ways on each of three dimensions of color, shape and motion (see Fig. 1 for examples). The training language for the first participant in each chain was a randomly generated set of 27 unique 2-4 syllable labels, built up from 9 possible CV syllables ('da', 'vi', 'ho', 'wi', 'nu', 'ri', 'bi', 'ka', 'tu'). These labels were randomly assigned to the 27 images, ensuring that there was a unique label for every image, with no systematic structure to the labels. The training language for later participants was the language produced by the previous participant in the chain during testing.

2.3 Procedure

2.3.1 Language learning, language testing, and dimension selection task

The participants worked through a computer program with three phases:

1) Learning phase.

   In each learning trial, the participant was presented with a label and two images, one of which was the target and one a distractor. The participant was instructed to pick which of the two images corresponded to the label. Once the

participant had clicked an image they were told whether their choice was correct or incorrect, shown the label and correct image for 2 seconds, and then instructed to retype the label before proceeding to the next trial. Target images were presented in random order. Distractors for each trial of the learning phase were assigned at random, subject to the following constraints: (i) within each learning block, each of the 27 meanings appeared once as a target and once as a distractor; (ii) according to the main experimental manipulation, one dimension was consistently backgrounded during learning and testing trials. For each participant, one of the three dimensions of shape, color and motion was selected as the backgrounded dimension. Every distractor then had the same feature as the target on this dimension (for example, if color was selected as the backgrounded dimension, the distractor on every trial would be the same color as the target). The other two dimensions were not manipulated in this way and served as controls. The learning phase of the experiment consisted of 4 blocks, each of 27 trials.

2) Test phase.

In each test trial, the participant was presented with two images: a target and a distractor. The target was highlighted with a black border. The participant was instructed to type the label that would let the alien know which image was highlighted. Target images were presented in random order. Distractors were randomly assigned within the same constraints as in the learning phase, i.e., they matched the target on the backgrounded dimension. The test phase consisted of 27 trials, one for each target.

3) Dimension selection task.

This final phase of the experiment used a method from Voiklis & Corter (2012) to test which dimensions participants thought essential to word meaning. On each trial, participants were presented with a label from the language they had been trained on and a concealed image. Their task was to decide whether the label-image pairing was correct or incorrect. In order to do this, they could click to reveal a feature of the concealed image (shape, color, motion), in any order. Participants could click Correct or Incorrect at any stage and did not have to reveal all features before doing so. A 1-second delay was included before features were revealed, to discourage participants from revealing features which were unnecessary to make the correct/incorrect judgment. The dimension selection task consisted of 27 trials, one for each image. Images were presented in random order. The labels for each trial were selected from the language the participant was trained on, such that 14 trials contained correct picture-label pairings and 13 incorrect picture-label pairings, but each label appeared only once.

2.3.2 Iteration

The language each participant produced in the test phase of the experiment was transformed and then used as the training language for the next participant in their chain. For this transformation, all dimensions and features of the images were randomly shuffled, so that patterns of labeling in relation to backgrounded and control dimensions were preserved, but individual correspondences of labels to images were

not (see Fig. 2 for an example). This transformation was intended to reduce the effects of intrinsic differences in salience of different dimensions, and to prevent the establishment of iconic labels (e.g., reduplicated syllables for bouncing images).

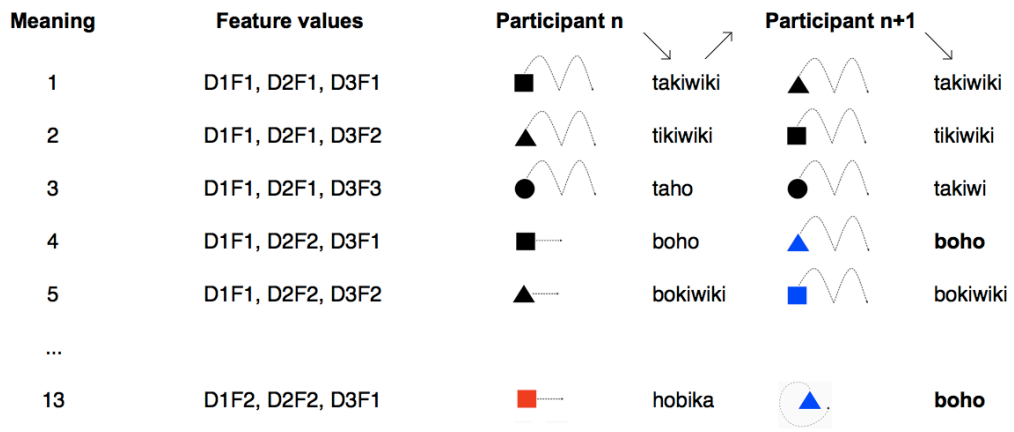


Figure 2. Illustration of the transformation process between participants in a chain. 'bk', black; 'rd', red; 'bl', blue; 'ci', circle; 'sq', square; 'tr', triangle; 'ho', moving horizontally; 'bo', bouncing; 'sp', spiraling. During the test phase, participant n produces mappings between 27 meanings (obtained from 3 features on Dimension D1 x 3 features on Dimension D2 x 3 features on Dimension D3) and 27 labels. The meaning of each label is therefore represented by specifying the Dimensions D1, D2, and D3 with features F1, F2, and F3 for each dimension. For example, for participant n, D1 is Color (where F1 = bk, F2 = rd, F3 = bl), D2 is Motion (where F1 = bo, F2 = ho, F3 = sp), and D3 is Shape (where F1 = sq, F2 = tr, F3 = ci). D1 is the backgrounded dimension (here, Color). The labels produced by participant n are presented to participant n+1 during the training phase; however, their corresponding meanings (i.e., the pictures) are changed randomly (while preserving the backgrounded and control dimensions). In the example, for participant n+1, D1 is Motion (where F1 = bo , F2 = sp, F3 = ho), D2 is Color (where F1 = bk , F2 = bl , F3

= rd), and D3 is Shape (where F1 = tr, F2 = sq, F3 = ci). The backgrounded dimension is still D1, but is now Motion rather than Color. The final column shows the new label produced by participant n+1 during their test phase. Here, we can see that while for participant n 'boho' means "black square moving horizontally", and 'hobika' means "red square moving horizontally", for participant n+1 'boho' means "blue triangle" regardless of motion. In other words, for meaning 13, this participant produces 'boho' where they were trained on 'hobika', changing the language with this error to introduce underspecification across the backgrounded dimension (motion for this participant).

2.4 Dependent variables

We used Kirby et al.'s (2008) measure of transmission error (how much the language produced by a participant during testing differed from their training input) to test whether the languages became more learnable over generations. Normalized Levenshtein edit distance between corresponding labels in successive generations (e.g. 'taho' and 'takiwi' for meaning 3 in Fig. 2) was calculated by taking the minimum number of edits (insertions, deletions, or substitutions of a single character) needed to transform one label into another, and then dividing by the length of the longer label. These values were then averaged across the whole language to give one measure of error per participant. If this value decreases over generations, the language is becoming more learnable.

Our specific hypothesis was that the languages would evolve gradually to underspecify more on the backgrounded dimension than on the control dimensions. Three outcome measures were taken to assess whether this was happening.

In order to capture the extent to which a language made distinctions on each dimension, we calculated (for each participant's test output) (1) the average number of words the language used across the features on each dimension (possible values ranging between 1 and 3); (2) average normalized Levenshtein edit distance between these labels, to give a more fine-grained measure of label dissimilarity. Fig. 3 gives an example of how these measures were calculated.

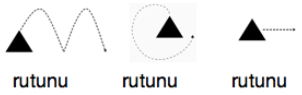
Meaning group	Number of words	Word dissimilarity
 nununu nununu nununu	1	0
 rutunu rutunu rutunu	1	0
 nununu ruhovu hovani	3	0.78
 rinunu rinununu rinununu	2	0.17
 rinunu rutunu rutunu	2	0.22
 rihonu rihova rihova	2	0.22
 danunu danunu danunu	1	0
 danunu danunu danununu	2	0.17
 dahovu dahovu dahovu	1	0
	1.67	**0.17**

Figure 3. Sets of meanings whose labels were compared to obtain the measures of language structure (here, with respect to the motion dimension). Meanings were divided into sets of 3 that differed only on one dimension. The word dissimilarity score is calculated by averaging the three normalized Levenshtein edit distances obtained by comparing the three possible word pairs. E.g., for row 4, rinunu/rinununu = distance 0.25, rinunu/rinununu = distance 0.25, rinunu/rinunu = distance 0, so the average word dissimilarity is 0.17. (Normalized Levenshtein edit distance for rinunu/rinununu: 2 letter additions necessary to turn one word into the other, divided by the length of the longest word (8) = 0.25.) Similar measurements are then made over all 9 sets of three meanings that differ only on the motion dimension, and these values averaged to give one underspecification value for that dimension for number of words, and one for word dissimilarity (values in bold).

Thirdly, participants' behavior on the dimension selection task (the order in which they chose to reveal the dimensions) was used to evaluate participants' attention to particular dimensions when evaluating word meaning. We gave a score of 3 for the dimension clicked first, 2 for second, 1 for third, and 0 if the dimension was not selected at all. Dimensions which are selected earlier, and are therefore presumably more central to word meaning, will have higher scores.

3. Results

Transmission error consistently decreased over generations (*M* generation 1 = 0.67, *SD* = 0.10; *M* generation 5 = 0.34, *SD* = 0.16). A linear trend ANOVA found that the trend was significant, $F(1,7) = 27.84$, $p < .001$, showing that the languages changed to become more learnable.

The results for the edit distance measure of underspecification are shown in Fig. 4, with the result for the dimension selection task in Fig. 5. Mixed-effects models were used for the main analyses of each of our dependent variables (number of words, within-dimension label dissimilarity, dimension selection task).[3] *p*-values for the fixed effects in these models were estimated using Baayen (2008)'s formula.[4] For post-hoc tests, the observations for the two control dimensions were averaged. Between-group *t*-tests were then run comparing backgrounded and control dimensions at each generation, applying the Bonferroni correction for multiple comparisons.

1) Number of words. Mean number of words across backgrounded and control dimensions was similar at generation 1 (*M* backgrounded = 2.93, *SD* = 0.10; *M* control = 2.91, *SD* = 0.13), then gradually diverged over generations 2-5, with more words remaining on control dimensions than backgrounded dimensions. The greatest difference was in generation 4: *M* backgrounded = 1.94, *SD* = 0.40; *M* control = 2.65, *SD* = 0.47). Fixed effects of dimension salience, generation, and an interaction were included in the mixed-effects model. Analysis of this model showed that the main effect of dimension salience was significant, $\beta = .25$, SE = .06, $t(144) = 4.46$, $p < .001$. There was also a significant linear trend for the number of words to decrease

[3] The random effects to include in these models were assessed by means of likelihood ratio tests. All models incorporated a random effect for Chain and a random slope for Participant.

[4] 2 * (1 - pt(abs(*t*), Y - Z)), where Y is the number of observations, and Z is the number of fixed effect parameters. The pt command on R accesses the probability distribution for *t*. Y-Z calculates the degrees of freedom, and multiplying by 2 obtains the *p*-value for a two-tailed test. Since this can be anticonservative at small sample sizes, we also used the heuristic of only accepting *t* values larger than 2 as significant (Baayen, 2008).

over generations, β = -.92, SE = .11, $t(144) = -8.29$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus control dimensions, β = .50, SE = .14, $t(144) = 3.66$, $p < .001$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of 0.008) found that the difference between backgrounded and control dimensions was marginally significant in generation 3, $t(7) = 3.54$, $p = .009$, and significant in generation 4, $t(7) = 4.03$, $p = .005$. The difference was not significant in any other generation ($t(7) < 1.71$, $p > .13$).

2) Within-dimension label dissimilarity (Fig. 4). Mean Levenshtein edit distance between words across backgrounded and control dimensions was similar at generation 1, then gradually diverged over generations 2-5. Words became more similar (i.e., edit distance was lower) on backgrounded dimensions than on control dimensions. The mixed-effects model incorporated main effects of dimension salience and generation, plus an interaction. Analysis of this model showed that the main effect of dimension salience was significant overall, β = .11, SE = .03, $t(144) = 4.30$, $p < .001$. Additionally, word dissimilarity tended to decrease over generations, β = -.40, SE = .05, $t(144) = -7.86$, $p < .001$, and the effect of generation was also significantly different for backgrounded versus control dimensions, β = .19, SE = .06, $t(144) = 2.97$, $p = .004$.

Post-hoc tests (using the Bonferroni correction to establish a significance criterion of .008) found that the difference between backgrounded and control dimensions was significant in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 4.42$, $p = .003$. The difference was not significant in any other generation ($t(7) < 1.95$, $p > .09$).

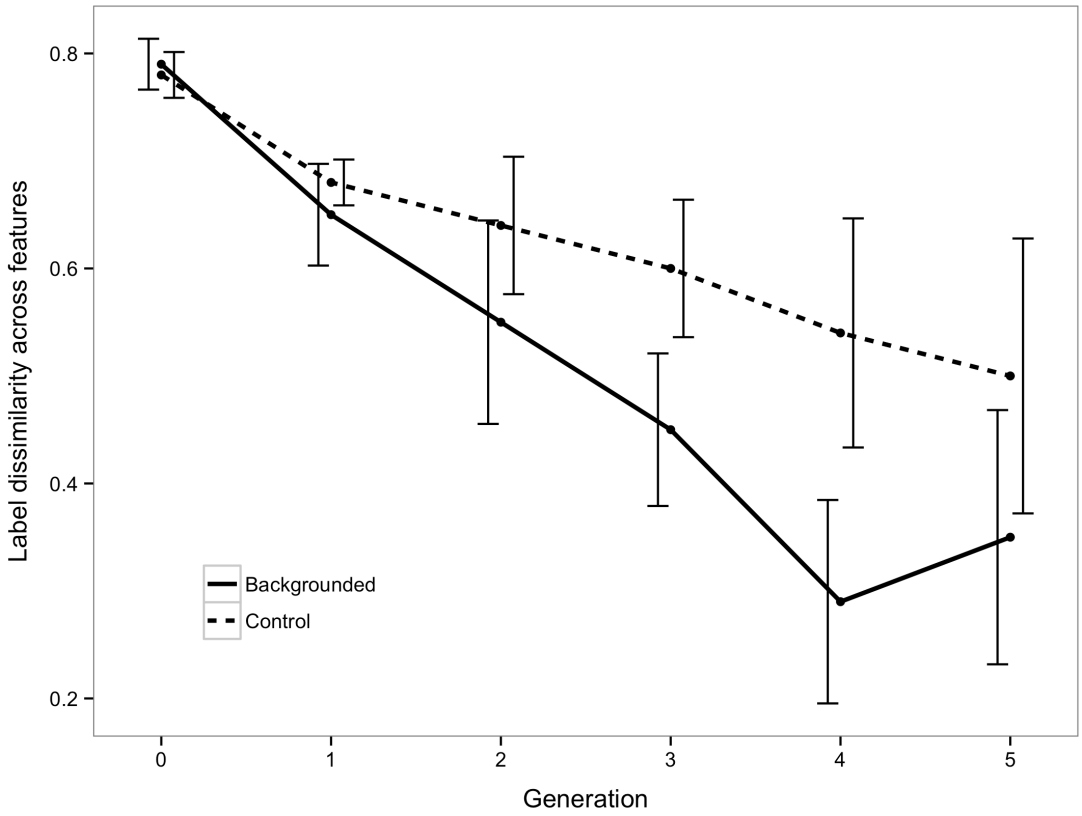


Figure 4. Dissimilarity of labels across features (see Fig. 3 for how this is calculated) against generation. The solid line indicates the backgrounded dimension, while the dashed line shows the control dimensions. Error bars (offset for clarity) show 95% confidence intervals. The results for number of words used across features were similar and are not shown (see text for descriptives).

3) Dimension selection task (Fig. 5). Mean selection preference score for backgrounded and control dimensions was similar at generations 1 and 2, then gradually diverged over generations 3-5, with higher preference scores on control dimensions than backgrounded dimensions. The mixed-effects model included fixed effects of dimension salience, generation, and an interaction. This model found

significant main effects of dimension salience (β = .45, SE = .12, $t(3240) = 3.85$, $p < .001$), generation (β = -.65, SE = .21, $t(3240) = -3.14$, $p = .002$), and a significant interaction between the two (β = .63, SE = .26, $t(3240) = 2.42$, $p = .016$).

Post-hoc tests found a significant difference between backgrounded and control dimensions only in generations 3, $t(7) = 3.92$, $p = .006$, and 4, $t(7) = 3.70$, $p = .008$ (significance criterion using Bonferroni correction = .01). The difference was not significant at any other generation, $t(7) < 1.09$, $p > .11$.

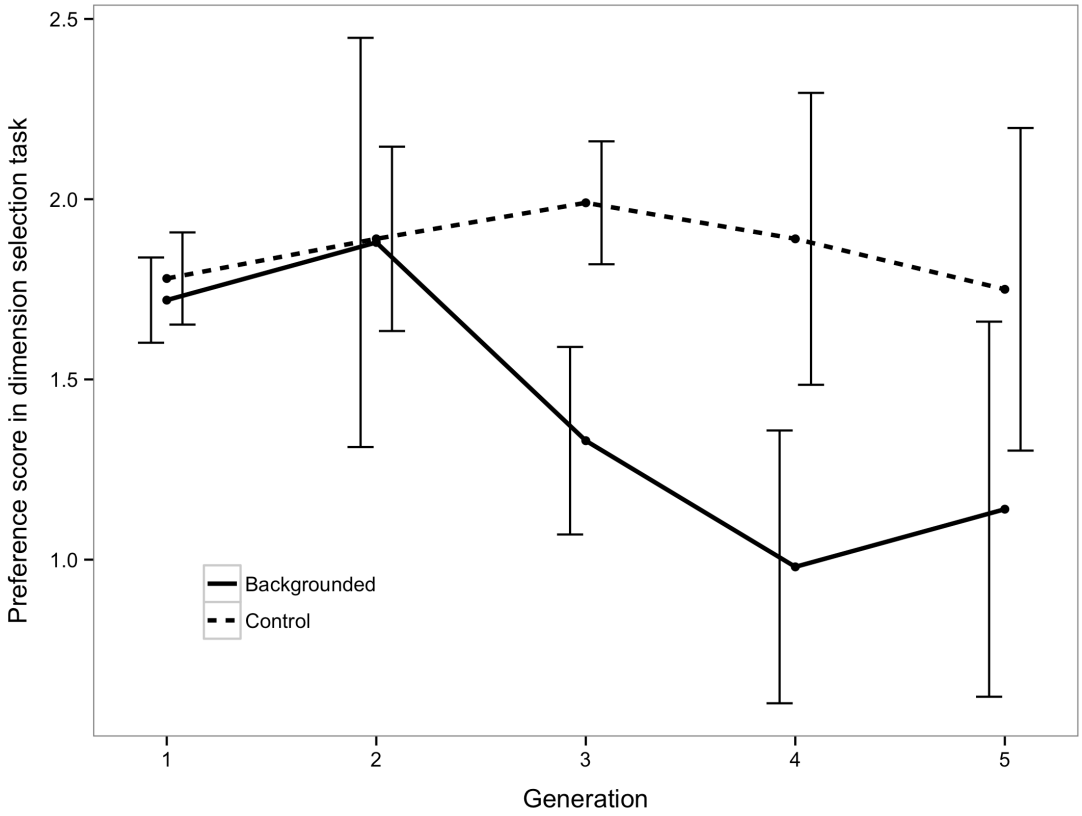


Figure 5. Change in attention to different dimensions over generations, evaluated via the dimension selection task. The solid line indicates the backgrounded dimension, while the dashed line indicates the control dimensions. Error bars are 95% confidence intervals.

Discussion

As predicted, patterns of underspecification that reflected the salience of dimensions in learning and production contexts arose gradually over generations of cultural transmission. Starting from input languages that specified equally across all dimensions, the languages lost distinctions earlier and faster on the dimension that was consistently backgrounded during learning and use. Fig. 6 shows a generation 5 language that underspecified more consistently on the backgrounded dimension (here, motion) than on the control dimensions. This was typical of the final languages in the experiment.

The gradualness of the effect is a product of individual-level learning processes amplified by cultural transmission. The first participant in each chain learns a language that sends a strong signal that all distinctions on all dimensions are important (since each image is labeled by a unique word). The performance of these participants on the dimension selection task shows that they have absorbed this expectation: they select all dimensions equivalently, showing that they consider them equally important to word meaning. However, this 27-word language is not learnable within the constraints of the training regime. Therefore, when these participants have to reproduce the language in the test phase, they are frequently faced with situations where they do not recall the word for the target referent. In this situation, a sensible strategy is to reuse a word they remember to be associated with at least one of the features of the referent. The question is, which feature(s) will they choose?

Globally, the initial language treats all dimensions as equally important. However, when participants are actually learning the meaning of each word, attending to the backgrounded dimension will never improve their success in the discrimination game.

This systematic manipulation means that the learner will tend to associate words more reliably with their referents' features on the more salient control dimensions than on the less salient backgrounded dimension. The analogous systematic structure of the production task, where the participant is cued to produce a word that will successfully discriminate the target from the distractor, also influences them to use a word which they associate with a feature on the salient dimension(s), rather than on the backgrounded dimension.

Therefore, participants will tend to reuse words for multiple referents that differ on the backgrounded dimension. The participant's task is still to converge on the language they are trained on, so errors in this direction will tend to be small and not necessarily systematic. However, as these errors build up over generations, they change the language and hence introduce a new source of evidence for the unimportance of one dimension: the patterns of word use in the language itself. Once a learner observes that a word can generalize over features on a dimension, this encourages the learner to reuse it for other cases if their memory fails (see Fig. 6 for a generation-by-generation view of how underspecification on the backgrounded dimension spreads as a chain progresses).

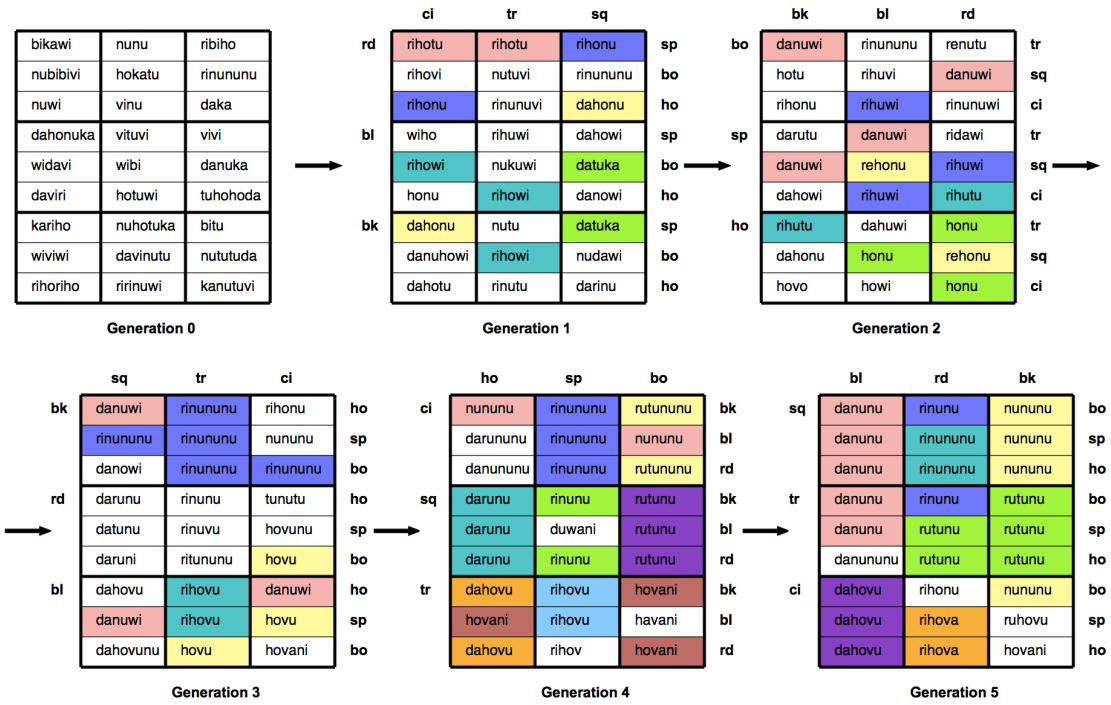


Figure 6. The emergence of underspecification in chain 7. Each grid shows one participant's language, arranged so the backgrounded dimension always runs down the right-hand side. Abbreviations as in legend to Fig. 2. Words used for more than one referent, i.e. underspecified words, are filled with the same color. The thick gridlines indicate regions that would be filled with the same color if the language underspecified on the backgrounded dimension. The figure shows underspecification arising more consistently on the backgrounded dimension than on either of the control dimensions, although it also extends partially to the control dimensions (see e.g. the 'overspill' of pink and green regions in generation 5).

The majority of the output languages in generation 5 were underspecified across more than just the backgrounded dimension. The generation 5 language in Fig. 6 underspecifies not only across the motion dimension (the backgrounded dimension

for this participant) but also partially across the shape dimension – e.g., a blue square and a blue triangle are both called ‘danunu’. This shows that the undirected underspecification that arose in Kirby et al. (2008) also occurred in this experiment, in addition to the underspecification cued by the experimental manipulation. The learning- and testing-based cues in this experiment, while sufficient to direct underspecification preferentially toward backgrounded dimensions, are not sufficient to prevent it eventually arising on control dimensions. This expected outcome of iterated learning leads to the lack of a significant difference between backgrounded and control dimensions in generation 5, as detailed in the Results. In real language use, other pressures presumably prevent undirected underspecification from happening, for example a pressure for unambiguous communication. One avenue for future work is to explore whether, with the introduction of a pressure for unambiguous communication (following Smith, Tamariz, & Kirby, 2013), underspecification would still emerge on the backgrounded dimension, while distinctions on control dimensions would be preserved.

Conclusion

We set out to investigate how patterns of underspecification that help learners generalize words appropriately could arise in language. The results show that attentional learning effects amplified over cultural transmission lead to a lexicon that underspecifies preferentially across dimensions that are habitually less salient during learning and use. Thinking of the language in the experiment as analogous to a particular region of the lexicon, for example object or substance nouns, illuminates a possible mechanism for the origin of the strong tendencies to specify on particular

dimensions we see in these regions. Over a whole language, specification will range over various dimensions depending on the function of individual words, as well as the characteristic situations in which they are learned and used. For example, the relational nature of gradable adjectives such as 'big' means that the contexts in which they are learned and used will tend to highlight dimensions of relations between objects as well as intrinsic dimensions (Clark & Amaral, 2010; Gentner & Kurtz, 2005; Sandhofer & Smith, 2001). More broadly, a language can be seen as a dynamic system where the meanings of individual words adapt to, as well as themselves contributing to, the salience of particular dimensions in contexts of learning and use. This result uncovers a mechanism for how words can come to specify in adaptive ways: as a cumulative product of the incremental changes made by individual learners attending to contextual cues in learning and use.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948-957.

Caldwell, C.A., & Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLOS ONE*, *7*(8), e43807.

Clark, E. V. (1993). *The lexicon in acquisition*. Cambridge: Cambridge University Press.

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, *64*(1), 1-37.

Clark, E. V, & Amaral, P. M. (2010). Children build on pragmatic information in language acquisition. *Language and Linguistics Compass*, *4*(7), 445-457.

Cornish, H. (2011). *Language adapts: Exploring the cultural dynamics of iterated learning*. (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh, UK.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351-86.

Geeraerts, D. (2009). *Theories of lexical semantics*. Oxford: Oxford University Press.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin* (pp. 151–175). Washington, D.C.: American Psychological Association.

Keil, F. C. (1994). Explanation, association, and the acquisition of word meaning. *Lingua*, *92*, 169-196.

Kelemen, D., & Bloom, P. (1994). Domain-specific knowledge in simple categorization tasks. *Psychonomic Bulletin & Review*, *1*(3), 390-395.

Kemler Nelson, D. (1995). Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts. *Cognitive Development*, *10*(3), 347-380.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(31), 10681-10686.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1153-1169.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199-227.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Perfors, A., & Navarro, D. (2011). Language evolution is shaped by the structure of the world: An iterated learning analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 477-482). Austin, TX: Cognitive Science Society.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317-328.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.

Sandhofer, C. M., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General*, *130*(4), 600-617.

Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually-cued attention and early word learning. *Cognitive Science*, *34*, 1287-314.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13-9.

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339-1343.

Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1348-1353). Austin, TX: Cognitive Science Society.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444-449.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308-312.

Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 956-961). Austin, TX: Cognitive Science Society.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401-413.

Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. *Cognitive Science*, *36*(4), 607-634.

Waxman, S. R., & Kosowski, T. D. (1990). Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. *Child Development*, *61*, 1461–1473.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–72.