https://doi.org/10.20378/irbo-50499

This is the post-print version of:
Tinits, Peeter, Jonas Nölle & Stefan Hartmann. 2017. Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution* 2(2). 148–159.
Link to publisher's version: https://doi.org/10.1093/jole/lzx011
Page numbers in parentheses show the page numbers in the final print.

Usage context influences the evolution of overspecification in iterated learning

Abstract

This paper investigates the influence of contextual pressures on the evolution of overspecification, i.e. the degree to which communicatively irrelevant meaning dimensions are specified, in an iterated learning setup. To this end, we combine two lines of research: In artificial language learning studies, it has been shown that (miniature) languages adapt to their contexts of use. In experimental pragmatics, it has been shown that referential overspecification in natural language is more likely to occur in contexts in which the communicatively relevant feature dimensions are harder to discern. We test whether similar functional pressures can promote the cumulative growth of referential overspecification in iterated artificial language learning.

Participants were trained on an artificial language which they then used to refer to objects. The output of each participant was used as input for the next participant. The initial language was designed such that it did not show any overspecification, but it allowed for overspecification to emerge in 16 out of 32 usage contexts. Between conditions, we manipulated the referential context in which the target items appear, so that the relative visuo-spatial complexity of the scene would make the communicatively relevant feature dimensions more difficult to discern in one of them.

The artificial languages became overspecified more quickly and to a significantly higher degree in this condition, indicating that the trend towards overspecification was stronger in these contexts, as suggested by experimental pragmatics research. These results add further support to the hypothesis that linguistic conventions can be partly determined by usage context and shows that experimental pragmatics can be fruitfully combined with artificial language learning to offer valuable insights into the mechanisms involved in the evolution of linguistic phenomena.

Keywords: overspecification, iterated learning, linguistic niche hypothesis, referential context, linguistic adaptation

1. Introduction

(p. 148) Natural languages often require certain meaning dimensions to be specified regardless of their relevance in the given communicative situation (Jakobson 1959, Dahl 2004). For example, some languages grammatically mark semantic features like animacy and shape, which are arguably redundant in most contexts (cf. McWhorter 2007). For example, the Austronesian language Nêlêmwa uses an animacy ergative marker to mark agents as animate (e.g., a person eating) or inanimate (e.g., water flowing; see ex. **(p. 149)** (1), Bril 2002: 158, 136). In these cases, the lexical referent alone would be enough to infer animacy . The overt conventional marking of communicatively redundant semantic distinctions has been termed grammatical overspecification (McWhorter 2012, Szmrecsanyi & Kortmann 2012).

(1)	a.	*Kio*	*i*	*khuxi*	*a*	*Pwayili*
		NEG	3SG	eat.TR	ERG.AN	Pwayili
		'Pwayili didn't eat it'				
	b.	T*axa*	*daan*	*ru*	*wi*	
		dig	road	ERG.INAN	water	
		'The water made holes in the road'				

Languages vary in the degree to which they exhibit grammatical overspecification, which has been identified as a key dimension of linguistic complexity (see e.g. McWhorter 2007, Lupyan & Dale 2010). Recent empirical research has frequently addressed potential factors that could lead to a loss in complexity (e.g. Lupyan & Dale 2010, 2016, Bentz & Winter 2013). While these studies aim to explain part of the cross-linguistic variation in complexity, they don't usually address the question of how overspecification can emerge in the first place (cf. Lupyan & Dale 2010, Trudgill 2011).

Increase in overspecification is often attributed to an accumulation of arbitrary changes that are maintained by the speakers (e.g. Dahl 2004: 103-118, 289-296; McWhorter 2007: 10-15, 21-28). However, recent research suggests that the development of linguistic complexity may be biased by environmental, pragmatic or sociocultural factors (e.g. Lupyan & Dale 2010, Trudgill 2011, Nettle 2012). According to the so-called Linguistic Niche Hypothesis, languages are shaped by the contexts in which they are learned and used and that differences between and within languages can partly be explained in these terms (Lupyan & Dale 2010, 2016; Dediu et al. 2013, Winters et al. 2015).

Experiments with artificial languages and emerging communication systems have been used to test specific hypotheses about such biases (Tamariz, 2017). Using non-conventional means of communication, allows for disconnecting the experimental task from participants' linguistic background knowledge. As such, the mechanisms involved in the evolution of communicative conventions can be addressed more directly. One factor that has repeatedly been addressed in these studies is referential context, which has been shown to systematically bias the evolution of languages towards systematic underspecification or compositionality (Silvey et al. 2015, Winters et al. 2015). This suggests that referential context may also play a role in the emergence of overspecification.

The contextual sensitivity of referential overspecification in natural language (e.g. using *Give me the red ball* when just one ball is in view; Arts et al. 2011: 555) has been investigated extensively in experimental pragmatics that employs artificial situations to identify conventions of use in particular languages. Given tasks where they have to refer to target items from an array, speakers tend to overspecify more frequently when the referential context has more items in it (Pechmann 1989, Koolen et al. 2011), or when the referential context is cluttered and difficult to perceive (Davies & Katsos 2009, 2013; Koolen et al. 2013, 2016). In these contexts, the relevant referential context is arguably harder to discern, making the relevant reference dimensions more difficult to find. In general, studies in natural language pragmatics indicate that speakers tend to overspecify certain meaning dimensions more when their communicative relevance is difficult to establish in the current context (Pechmann 1989, Koolen et al. 2016). From a diachronic perspective, this leads to the question whether this contextual variation can serve as an evolutionary mechanism influencing the frequency of referential overspecification in a language over time.

The increasing use of referential overspecification can be seen as a major factor giving rise to phenomena that have been identified as instances of grammatical overspecification. Bybee et al. (1995: 8), for example, mention frequency increase and contextually redundant usage as factors that conspire in the process of grammaticalization. As specific signs are used more frequently and extend to new contexts, they can become obligatory and thus give rise to grammatical overspecification (Heine & Kuteva 2007; Lehmann 2015). Arguably, identifying biases that influence the evolution of referential overspecification can thus provide insights as to the development of grammatical overspecification. If the aforementioned contextual variation can influence the frequency of referential overspecification over time, real-world circumstances that offer different types of contexts could also make the emergence or conventionalization of overspecification more likely. In line with the Linguistic Niche Hypothesis, this could in turn help explain cross-linguistic differences in grammatical overspecification.

The present study combines artificial language dynamics and experimental pragmatics to investigate the role of context in the evolution of overspecification. Following earlier studies that investigate the effect of referential context (e.g. Silvey et al. 2015, Winters et al. 2015), we apply the iterated learning paradigm (Kirby et al. 2008), in which participants are trained on an artificial language and their output is passed on to **(p. 150)** the next "generation" of participants. The reuse of participants' outputs allows us to monitor how the effects of context biases accumulate over time. This can help identify states on which the learning materials converge on, indicating potential "cultural attractors", i.e. states that are more likely to be attained and preserved during cultural transmission (Mesoudi & Whiten 2008, cf. Sperber 1996). Such biases can potentially explain general patterns in linguistic structure.

Our experiment uses this approach to test whether the contextual pressures suggested by pragmatic studies can also promote the cumulative growth of referential overspecification in the cultural transmission of an artificial language. Such contextual pressures, as argued above, can presumably shape the diachronic development of grammatical overspecification in natural languages. We used two conditions, which differed in terms of visuo-spatial complexity. The experiment required participants to learn an artificial language – with initially no overspecification present – and use it to solve a referential task in a context that allowed for overspecification to be introduced. We predicted that the difference between conditions would lead to a bias towards more overspecification in the more complex context in the evolution of the miniature language.

2. Methods

2.1 Participants

205 volunteers were recruited online and allocated into 41 chains of 5 individuals each. Participants received no compensation. Five chains were excluded from analysis due to noncompliance with the instructions (e.g. using English words and phrases like *nope*, *no*, *no idea*). In six other chains, the color markers were no longer reliably associable with colors at some point in the chain. Overspecification on this dimension was then no longer possible, which is why these chains were also omitted from the analysis. Details on this are included in Appendix 3.

2.2 Materials

Language. The participants were trained on a language that described four objects (*pen*, *ball*, *book*, *cup*) in two colours (*yellow*, *blue*) in various contexts. In all chains, the first participant in the chain was presented with a language that was "minimally specified" (cf. Pechmann 1989) on one dimension: it only specified the color dimension if it was necessary to disambiguate the target object from the distractor (see Tab. 1).

Tab. 1. The initial language. The meaning space consisted of four objects in two colors. In contexts that required disambiguation (middle and rightmost column), the object label was used with a color marker, e.g. *meeb li* 'yellow pen' or *meeb pu* 'blue pen'. In all other cases, only the object label was used, e.g. *meeb* 'pen' (left column).

Object	No con-trast		Blue marked		Yellow marked
	meeb		*meeb pu*		*meeb li*
	fop		*fop pu*		*fop li*
	kur		*kur pu*		*kur li*
	yan		*yan pu*		*yan li*

Visual stimuli. Participants were trained on the alien language by showing them images with accompanying labels. Each image showed one or two out of four everyday items (*pen*, *ball*, *book*, *cup*) either in yellow or in blue. The sets of objects presented on each trial differed across conditions (see Section 2.3 below).

Conditions. Participants were allocated to one of two conditions. Across conditions, 16 of the 32 trials required disambiguation between two objects of the same type (e.g. a blue cup and a yellow cup, see Fig. 1, left panel, bottom row) (control trials). In the **(p. 151) simple context condition**, only one single object was displayed in the remaining 16 trials (test trials). In the **complex context condition**, by contrast, the other 16 trials consisted of pictures showing two different objects (e.g. cup and pen). Both color (yellow vs. blue) and position (left vs. right) were assigned randomly to the objects in these pictures, such that the color of target and distractor would match in half of the trials (as in Fig. 1) and differ in the remaining trials, and the target would be on the left in half of the trials and on the right in the remaining trials. Contexts in which more than one feature dimension needs to be distinguished, such as recognizing the irrelevance of color in test trials of the complex condition, have been shown to be more costly in processing (Treisman & Gormican 1988). The similarity of test and control trials therefore makes the complex condition harder to process.

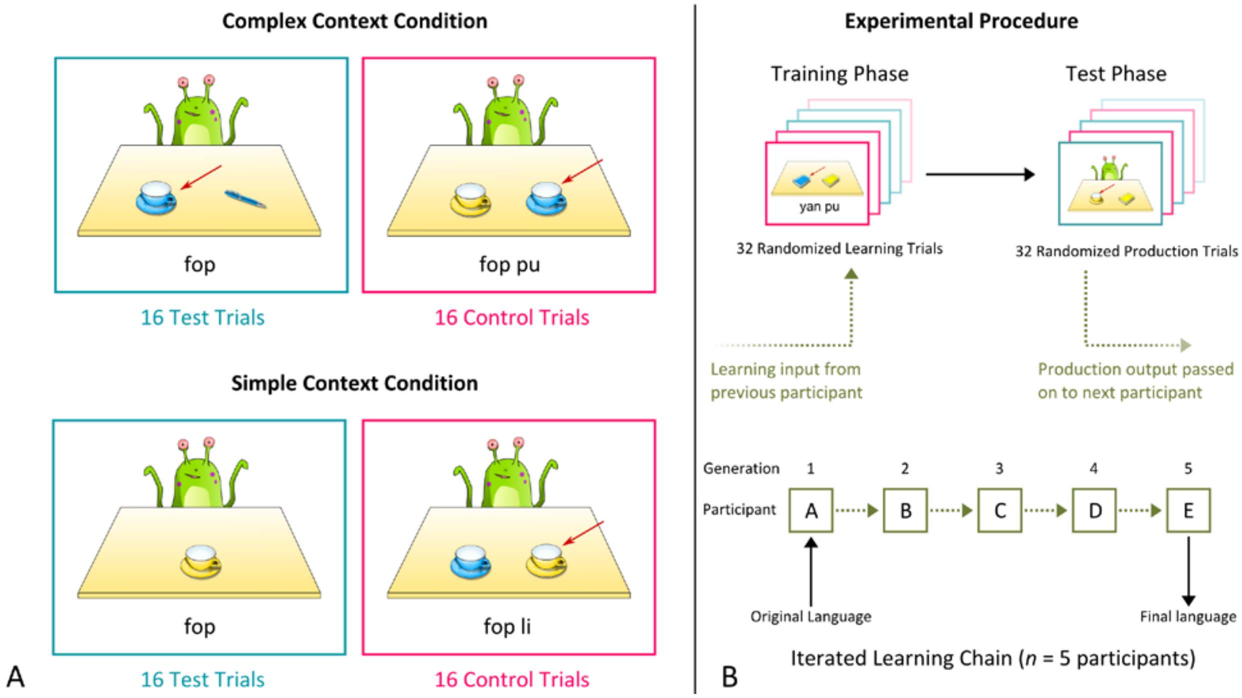


Fig. 1. Overview of the experimental setup. A) Test trials differed across conditions. In the simple context condition, one item was shown in isolation; in the complex context condition, two items of the same type, but with different colors, were shown. In these trials, no color markers were used in the initial language. Control trials that required color specification were identical across conditions. B) Each participant, being part of a chain of 5, had to learn the language and reproduce it. The output of participant n became the input for participant n+1.

2.3 Procedure

The participants were first provided a training sequence of 32 randomized trials, where they would "learn to communicate" with an alien to help it find the right object (see the instructions in Appendix 1). The training stimuli consisted of images accompanied by artificial language labels, which participants had to retype after seeing them for 2 seconds. No feedback was provided.

After the training sequence they were told to help the alien find the right object by using the alien language and provided the same 32 trials in a random sequence. They were explicitly asked not to use any other language they know. The output was stored and given as learning input for the next participant in the chain. This was repeated four times in each chain, amounting to 5 "generations".

2.4 Analysis

To assess the degree of overspecification in each chain, the number of color markers in the 16 test trials was counted. Items were considered color markers if they reliably distinguished between blue and yellow in the control trials.

3. Results

30 chains entered the analysis (13 in simple, 17 in complex) amounting to 150 linguistic output sets and a total of 2400 potentially overspecified utterances. In both conditions, the experiment afforded the full continuum from minimal to maximal overspecification to be produced by the first generation of learners.

(p. 151) The output was transmitted between participants exactly as produced (unlike, e.g., Smith & Wonnacott 2010 who applied a manual filter), which introduced variation along other dimensions than color marking: In 18 chains, the language maintained distinct signals for all four objects and both colors across all generations. The remaining 12 chains introduced some noise: 1) in nine of them, an object label was lost at some point; 2) in three chains, the expansion of one of the object labels led to complete homonymy between two or more objects.

As this variation can potentially confound the evolution of overspecification, the analysis was performed on both the full dataset of 30 chains and the subset of 18 chains that remained fully expressive throughout (henceforth "expressive subset"). Across all 30 chains, the deviations from the word-forms in the learning input were marginal[1].

On average, both conditions showed a trend towards overspecification, but the complex condition to a greater extent (see Fig. 2). Altogether, the full dataset contains 472 overspecified color markers in the isolated condition as compared to 1050 in the complex condition; in the expressive subset, the proportion is 277:596.

[1] To measure signal fidelity over generations, we used normalized Levenshtein distance (Levenshtein 1966), calculated by dividing the minimal number of insertions, deletions and substitutions required to transform one label into another, by the length of the longer label. Comparing each word to its closest match in the training language (see e.g. Vokey & Brooks 1992, Cornish et al. 2017), it can be shown that the mean deviation per generation remains below 0.05 across all generations. See supplementary materials for more details.

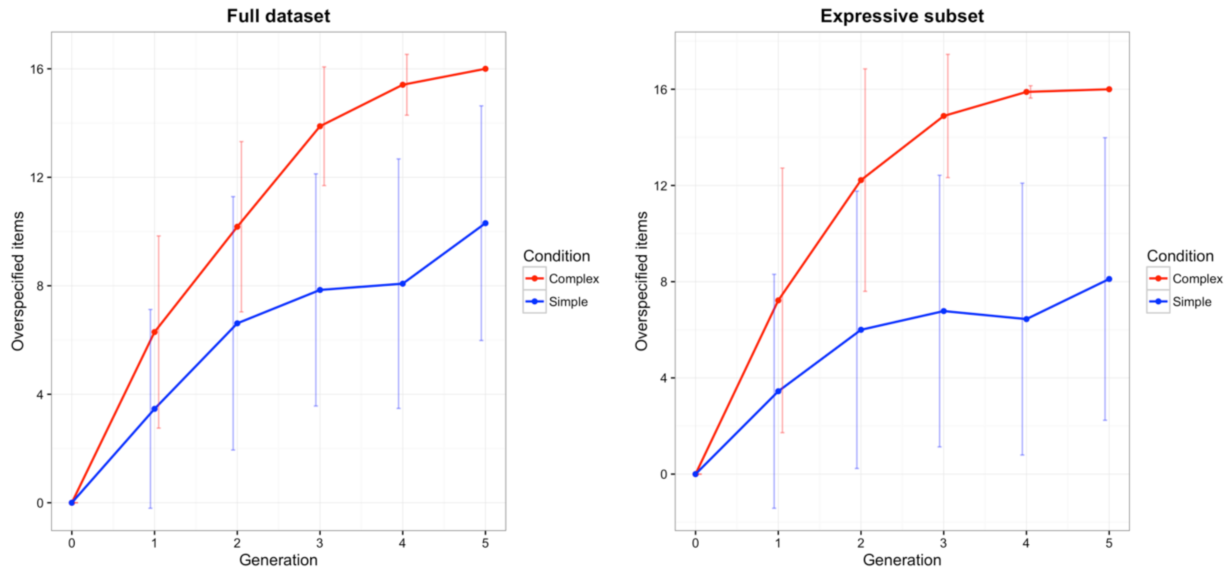


Fig. 2. Number of overspecified color markers in the full dataset (left panel) and the expressive subset (right panel). Error bars show 95% confidence intervals.

To test the influence of the experimental manipulation while controlling for other factors, a binomial growth curve model (see Mirman 2014; Winter & Wieling 2016) was fit to the data using R (R Core Team 2015) and lme4 (Bates et al. 2015). The use of an overspecific color marker at each individual datapoint is used as the (binary) response variable. The predictor variables are *Condition* (simple vs. complex) and *Generation*, whereby the latter is also added as a quadratic fixed effect. Native language and the chain to which the individual datapoint belongs were added as random effects (see Appendix 2 for a more detailed discussion of the model). According to a log-likelihood test, the model performs significantly better than a binomial mixed-effects regression model without the polynomially transformed predictor (full dataset: $\chi2=27.05$, $df=2$, $p<0.001$; expressive subset: $\chi2=18.38$, $df=2$, $p<0.001$). More importantly, it significantly outperforms null models without *Condition* when *Condition* is removed as predictor (full dataset: $\chi2=98.7$, $df=4$, $p<0.001$; expressive subset: $\chi2=110.83$, $df=4$, $p<0.001$) or without the interaction between *Condition* and *Generation* (full dataset: $\chi2=10.61$, $df=2$, $p<0.01$; expressive subset: $\chi2=11.7$, $df=2$, $p<0.01$). In sum, then, the model lends support to the **(p. 153)** hypothesis that overspecification was generally higher in the complex context condition and that the trend of *Generation* differs between conditions.

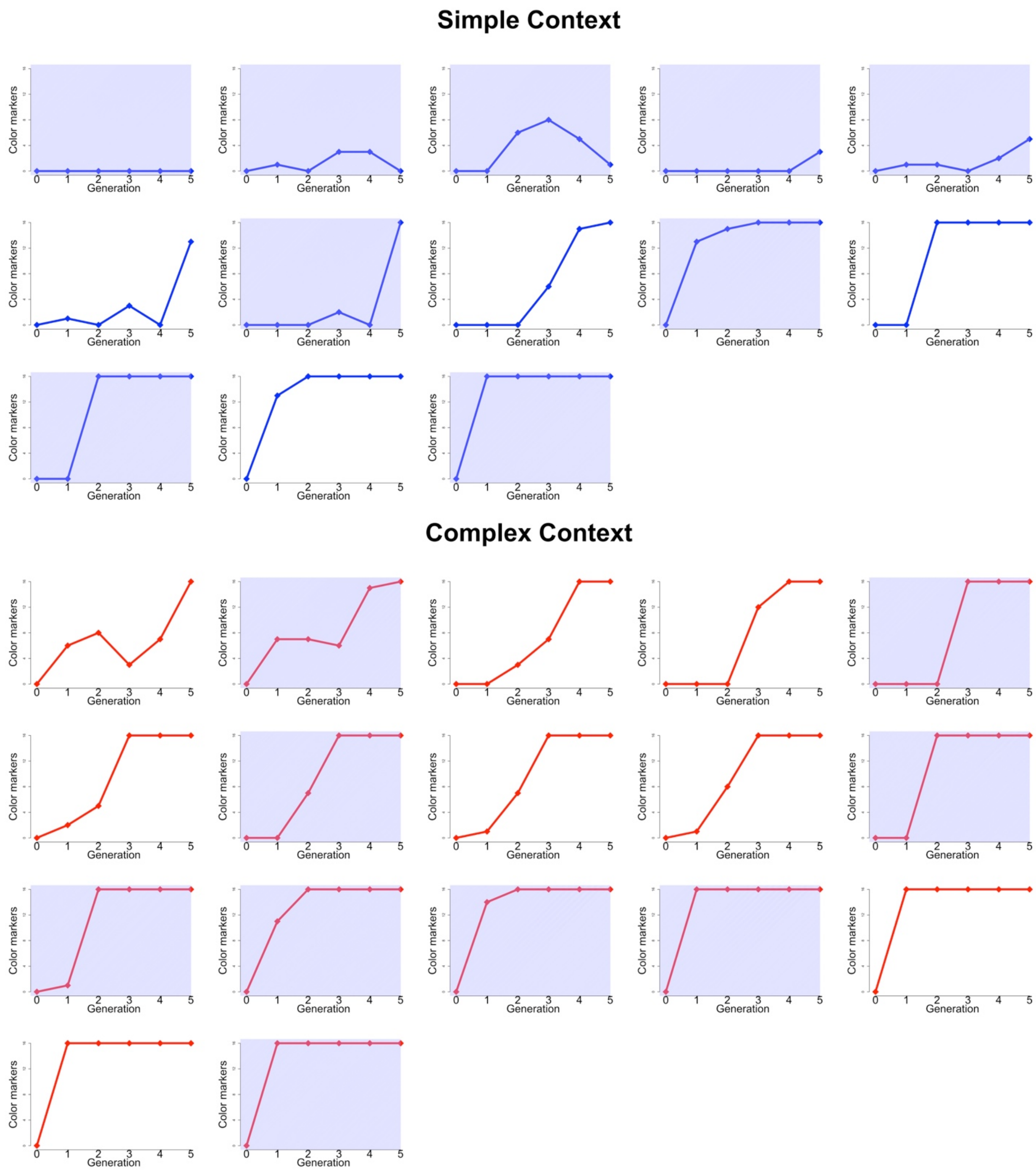


Fig. 3. Development of the individual chains in each condition. Shaded panels: expressive subset.

Inspection of individual chains (Fig. 3) reveals that similar changes tended to happen asynchronically across conditions. For instance, on 10 occasions, a minimally specified language was made maximally overspecified by a single learner at different times in the chain. Previous work has demonstrated tendencies for artificial language learning to develop towards less free variation (Smith & Wonnacott 2010, Fehér et al. 2016, Roberts & Fedzechkina 2016). In the present experiment, the tendency went towards maximum overspecification: As soon as some overspecification was introduced to **(p. 154)** the language, it tended to increase further, and in all chains where the maximum was reached, it remained stable onward. This upward trend proved unidirectional as soon as at least 9 items were overspecified. However, in the simple condition, overspecification also frequently showed a decrease at some point in the chain (this was rare in the complex condition), sometimes even restoring the minimally specified system, indicating that the trend towards overspecification was weaker here.

4. Discussion

In sum, the results of our experiment demonstrate that contexts of use can significantly influence the evolution of referential overspecification in iterated learning: in the contexts where the relevant feature dimensions are harder to discern, the trend towards overspecification is stronger. These differences in overspecification can be attributed to an interaction between pressures towards overspecification and the tendency to reduce variation that is known from earlier studies: Both minimal and full overspecification seem to act as "attractor" states (see Fig. 3). When the general tendency to overspecify is weaker, the language can also return to the initial degree of specificity, which can also be interpreted as reduction in variation, as the language reverts to full systematicity by only specifying color markers when they are communicatively necessary.

These results can be seen to lend further support to the Linguistic Niche Hypothesis, as our experiment shows how the dynamics of language learning and use can differ due to contextual variation. While the experiment makes use of a highly artificial experimental manipulation, contexts that influence the discriminability of referents in natural language may include, factors such as the physical, sociocultural, or technological environment.

The differences observed in our artificial language learning study conformed to the predictions made on the basis of experimental pragmatics studies (e.g. Pechmann 1989, Koolen et al. 2016): a difficulty in establishing the relevance of a reference dimension can make overspecification relatively more likely to increase in iterated learning.

However, some caveats remain, such as the question whether the maximal use of overspecification in our experiment constitutes or at least paves the way towards full conventionalization and "obligatorification" (see Lehmann 2015), i.e. towards color marking being a structural feature of the resulting language. The present study can only detect biases involved in the growth of referential overspecification. However, as outlined in the Introduction, these can reasonably be assumed to play a role in the emergence of grammatical overspecification as well.

Further studies could extend the present findings and make up for some of these shortcomings by introducing more complex experimental setups to investigate the nature of the effect we found in more detail. For example, we compared simple linear transmission chains (Mesoudi & Whiten 2008), which allowed us to look at a bias due to contexts of use in two different settings (simple vs complex). Future studies could extend this by investigating a mixture of these contexts (Silvey et al. 2015, Winters et al. 2015) or manipulating their assumed common ground (Winters et al. 2016). Using these paradigms would enable testing to what degree the emerging redundant use of semantic markers in the miniature language is actually becoming obligatory regardless of context, and allow for the participants' expectations as to what is actually communicatively necessary to take into account.

Also, we focused on the dimension of color, which has been found to be often overspecified in earlier studies (Pechmann 1989, Gatt et al. 2011, Koolen et al. 2013). At the same time, the development of obligatory color markers is so far unattested in natural languages (Talmy 2000: 24). Further studies could look at various types of referents and meaning dimensions (analogous to, for example, tense, aspect, gender or number in natural languages; cf. e.g. Culbertson et al. 2009, Hudson Kam & Newport 2009, Fedzechkina et al. 2011, Tily et al. 2011). Also, structural parameters like word-order could be manipulated (e.g. Rubio Fernández 2016 found color overspecification to be weaker in a language with postnominal adjectives). Taking into account different types of referents and structural parameters of the starting language would allow for checking if the results obtained may have been due to the specific experimental setup used here.

As the present study draws on a pseudo-communicative setup in which participants point out objects to a depicted, unreactive alien, future studies could include various communicative settings, e.g. cooperative setups with co-present individuals who can react to each other. Previous studies indicate that imaginary addressees might increase the level of overspecification (see van der Wege 2009), while the presence of another speaker tends to keep the level of specificity minimal (see Clark & Wilkes-Gibbs 1986, Arts et al. 2011). It is worthwhile to explore how these tendencies interact with the dynamics of learning and use across generations.

To conclude, our experimental results show that context can significantly influence the evolution of overspecification in iterated learning. This adds support to the **(p. 155)** hypothesis that language conventions can be partly determined by usage context (cf. e.g. Winters et al. 2015). Using artificial language learning to test predictions motivated by studies from experimental pragmatics, our experiment shows that the trend towards overspecification is stronger in contexts in which the communicatively relevant feature dimensions are more difficult to discern, as predicted on the basis of previous studies on natural language use. As such, the experiment demonstrates how pressures that have been established in synchronic experimental studies, along with the dynamics of language learning and use, can give rise to significant differences in the development of miniature languages. These pressures may also play a role in the cultural evolution of conventional overspecification in natural languages, given that the dynamics of learning and transmission rely on analogous principles.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

References

Software

R Core Team. 2016. R. A software and environment for statistical computing. Vienna: R Foundation for statistical computing.

R packages:

Bates, Douglas et al.: lme4. Linear Mixed-Effects Models using 'Eigen' and S4.

Hothorn, Thorsten & Hornik, Kurt: exactRankTests 0.8-28. Exact Distributions for Rank and Permutation Tests.

Wickham, Hadley & Chang, Winston: ggplot2 2.1.0.

Research literature

Arts, Anja, Alfons Maes, Leo G. M. Noordman & Carel Jansen. 2011. Overspecification in written instruction. *Linguistics* 49(3). 555–574.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). doi:10.18637/jss.v067.i01.

Bentz, Christian, & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3(1). 1-27.

Bril, Isabelle. 1996. Split ergativity in the Nêlêmwâ language. In Cecilia Odé & Wim Stokhof (eds.), *Proceedings of the Seventh International Conferece on Austronesian Languages*, 377–393. Amsterdam, Atlanta: Rodopi.

Bybee, Joan L., Revere Perkins & William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Clark, Herbert H & Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22(1). 1–39.

Cornish Hannah, Rick Dale, Simon Kirby, Morten H Christiansen 2017. Sequence Memory Constraints Give Rise to Language-Like Structure through Iterated Learning. *PLoS ONE* 12(1). e0168532.

Culbertson, Jennifer, Paul Smolensky & Géraldine Legendre. 2009. Testing Greenberg's Universal 18 using the Mixture Shift paradigm for artificial language learning. *Proceedings of NELS*, vol. 40.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam, Philadelphia: John Benjamins.

Davies, Catherine & Napoleon Katsos. 2009. Are Interlocutors as Sensitive to Over-informativeness as they are to Under-informativeness? *Proceedings of the Workshop on Production of Referring Expressions: Bridging Computational and Psycholinguistic Approaches*. http://pre2009.uvt.nl/pdf/davies_katsos_7_8.pdf.

Davies, Catherine & Napoleon Katsos. 2013. Are speakers and listeners "only moderately Gricean"? An empirical response to Engelhardt et al.(2006). *Journal of Pragmatics* 49(1). 78–106.

Fedzechkina, Mariya, T. Florian Jaeger & Elissa Newport. 2011. Functional biases in language learning: Evidence from word order and case-marking interaction. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33th Annual Conference of the Cognitive Science Society*, 318–323. Austin: Cognitive Science Society.

Fehér, Olga, Kenny Smith, Nikolaus Ritt & Elizabeth Wonnacott. 2016. Communicative interaction leads to the elimination of unpredictable variation. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Féher & Tessa Verhoef (eds.), *The Evolution of Language. Proceedings of the 11th International Conference*, 415–416.

Gatt, Albert, Roger van Gompel, Emiel Krahmer & Kees ven Deemter. 2011. Non-deterministic attribute selection in reference production. *Proceedings of the Workshop on Production of Referring Expressions: Bridging the gap between empirical, computational and psycholinguistic approaches to reference (PRE-CogSci'11).* https://pre2011.uvt.nl/pdf/precogsci2011.pdf.

Heine, Bernd & Tania Kuteva. 2007. *The Genesis of Grammar: A Reconstruction*. Oxford: Oxford University Press.

Hudson Kam, Carla L & Elissa L Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive psychology* 59(1). 30–66.

Jakobson, Roman. 1959. On linguistic aspects of translation. In Reuben A. Brewer (ed.), *On translation*, 232–239. Cambridge: Harvard University Press.

Kirby, Simon, Hannah Cornish & Kenny Smith. 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* 105(31). National Acad Sciences. 10681–10686.

Koolen, Ruud, Albert Gatt, Martijn Goudbeek & Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics* 43(13). Elsevier. 3231–3250.

Koolen, Ruud, Martijn Goudbeek & Emiel Krahmer. 2013. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science* 37(2). Wiley Online Library. 395–411.

Koolen, Ruud, Emiel Krahmer & Marc Swerts. 2016. How Distractor Objects Trigger Referential Overspecification: Testing the Effects of Visual Clutter and Distractor Distance. *Cognitive Science* 40(7). 1617–1647.

Lehmann, Christian. 2015. *Thoughts on Grammaticalization*. 3rd ed. (Classics in Linguistics 1). Berlin: Language Science Press.

Levenshtein, Vladimir I 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. 10(8). 707–710.

Lupyan, Gary & Rick Dale. 2010. Language Structure is Partly Determined by Social Structure. *PLoS One* 5(1). doi:10.1371/journal.pone.0008559.

Lupyan, Gary & Rick Dale. 2016. Why are there different languages? The role of adaptation in linguistic diversity. Trends in Cognitive Sciences. 20(9). 649–660 DOI: http://dx.doi.org/10.1016/j.tics.2016.07.005

McWhorter, John. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press.

McWhorter, John. 2012. Complexity hotspot. The copula in Saramaccan and its implications. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic complexity. Second language acquisition, indigenization, contact*, 243–264. (Linguae & Litterae 13). Berlin, Boston: de Gruyter.

Mesoudi, Alex & Andrew Whiten. 2008. The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363(1509). 3489–3501.

Mirman, Daniel. *Growth curve analysis and visualization using R*. Boca Raton: CRC Press.

Nettle, Daniel. 2012. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367(1597). 1829–1836.

Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics* 27(1). 89–110.

Roberts, Gareth & Mariya Fedzechkina. 2016. Social biases versus efficient communication: An iterated learning study. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Fehér & Tessa Verhoef (eds.), *The Evolution of Language. Proceedings of the 11th International Conference*, 526–527.

Rubio-Fernández, Paula. 2016. How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology* 7. doi:10.3389/fpsyg.2016.00153.

Silvey, Catriona, Simon Kirby & Kenny Smith. 2015. Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive science* 39(1). 212–226.

Smith, Kenny & Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. *Cognition* 116(3). 444–449.

Sperber, Dan. 1996. *Explaining culture: A naturalistic approach*. Cambridge Univ Press.

Szmrecsanyi, Benedikt & Bernd Kortmann. 2012. Introduction: Linguistic complexity. In Bernd Kortmann & Benedikt Szmrecsanyi (eds.), *Linguistic complexity. Second language acquisition, indigenization, contact*, 6–34. (Linguae & Litterae 13). Berlin, Boston: de Gruyter.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Vol. 1. Cambridge, Mass: MIT Press.

Tamariz, Monica. 2017. Experimental Studies on the Cultural Evolution of Language. *Annual Review of Linguistics* 3(1). 389–407.

Tily, Harry, Michael C Frank & T Florian Jaeger. 2011. The learnability of constructed languages reflects typological patterns. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science*, 1364–1369. Austin: Cognitive Science Society.

Treisman, Anne & Stephen Gormican. 1988. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95(1). 15–48.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

van der Wege, Mija M. 2009. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language* 60(4). 448–463.

Vokey, John R, Lee R Brooks 1992. Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 18(2). 328–344.

Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution* 1(1). 7–18.

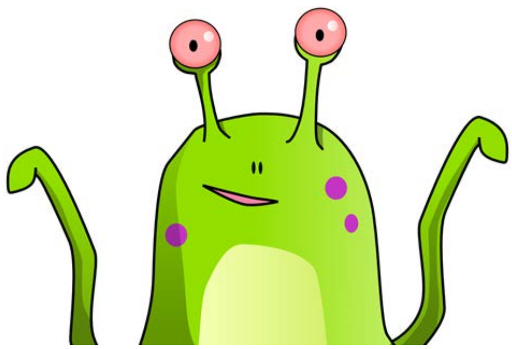
Winters, James, Simon Kirby & Kenny Smith. 2015. Languages adapt to their contextual niche. *Language and Cognition* 7(3). 415–449.

Winters, James, Simon Kirby & Kenny Smith. 2016. Signal autonomy is shaped by contextual predictability. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Olga Fehér & Tessa Verhoef (eds.), *The Evolution of Language. Proceedings of the 11th International Conference*, 605–606.

Appendix 1. Experimental setup

Instructions

Training phase



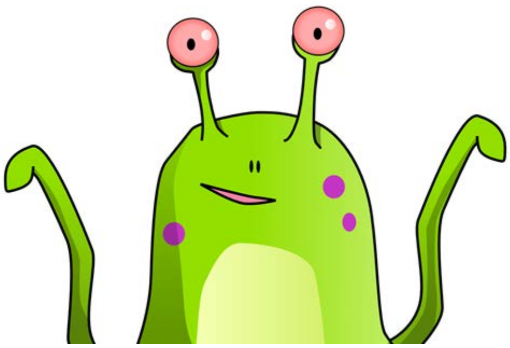
Welcome to our little experiment! It will take about 15 minutes.

This is Ari, an alien from the planet Tsamtrah. Over the next few minutes you will learn to communicate with Ari. You are dealing with a shipment of earth items to Ari which got mixed up on space travel. Ari has restored most of the order, but needs your help in the finishing steps.

Your task is to help Ari in various situations where Ari is unsure which object to pick from the table. In some cases Ari will just need your confirmation that the job has been done right. Your screen on earth indicates the right item to choose next. You will only have your keyboard to help with this task.

Press ENTER to continue

Communication phase



Congratulations - you have finished a crash course in Tsamtrahian!

You now get to use your new knowledge. Help Ari find the right object from the shipments he got from earth! You can use your keyboard to write just as you did before. Just press ENTER when done, and the signal will be transmitted to Ari.

If you’re not entirely sure about a word, please still try and enter your best guess in Tsamtrahian - please DON’T use English or any other language: Ari won’t understand you. This could result in misunderstanding and even war, which might mean the end of our civilization.

Good luck! Click the button below to continue.

Appendix 2. Methodology

Specifics of the Growth Curve Model

A growth curve model (cf. e.g. Winter & Wieling 2015) with the specification given in (1) was fit both to the full dataset and to the expressive subset. Two participants who did not specify their native language were omitted from the analysis. In the case of two bilingual participants, only the language they mentioned first was taken into consideration. The model specification is given in (1).

(1) OverspecificColorMarkers~Condition+Generation+Generation^2+
Condition:Generation+Condition:Generation^2+(1|nativeLanguage)+(1|Chain)

Table A2 shows the parameter estimates for the fixed effects of the model based on the full dataset along with their standard errors as well as p-values based on asymptotic Wald tests as provided by *lme4*.

Tab. A2. Model summary (fixed effects) for the growth curve model fit to the full dataset.

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	0.78	1.41	0.58
Condition_complex	3.94	1.68	0.02*
Generation	1.59	0.20	9.28E-16***
Generation²	-0.49	0.10	1.80E-06***
Condition_comp:Generation	0.48	0.27	0.08 .
Condition_comp:Generation²	0.41	0.15	0.007 **

Note: * p < .05, ** p < .01, *** p < .001

According to a log-likelihood test, the model performs significantly better than a binomial mixed-effects regression model without the polynomially transformed predictor (full dataset: χ^2=27.05, *df*=2, p<0.001; expressive subset: χ^2=18.38, *df*=2, p<0.001). More importantly, it significantly outperforms null models without *Condition* (full dataset: χ^2=98.7, *df*=4, p<0.001; expressive subset: χ^2=110.83, *df*=4, p<0.001) or without the interaction between *Condition* and *Generation* (full dataset: χ^2=10.61, *df*=2, p<0.01; expressive subset: χ^2=11.7, *df*=2, p<0.01). Note that the random effect of *Native Language* also significantly improves the model fit (full dataset: χ^2=78.1, *df*=1, p<0.001; expressive subset: χ^2=89.1, *df*=1, p<0.001). However, it is unclear to what degree this factor actually has an effect and to what degree this result might be an artefact of the rather imbalanced composition of the participants' native languages: Altogether, the participants have 21 different native languages, but no less than 58 are native speakers of German, 31 of English, and 17 of Estonian, while all other languages are represented with less than 10 speakers. One obvious possibility is that the order of adjective and noun in the respective native languages might have an effect: Our stimuli bear obvious resemblance to noun-adjective patterns in natural languages. Thus, they might be learned differently by speakers whose native language features noun-adjective word ordering. However, this is not the case as the nine participants who speak a noun-adjective language (according to Dryer 2013) cover the whole span from minimal specificity to full overspecification.

Functionality of the color markers

Six chains were excluded from the analysis as the color markers became dysfunctional at some point in the chain. In some cases, this was due to the complete loss of one of the markers; in other cases, they were used as lexical markers that associated with particular objects or all objects or used in a way that made them unlikely to be understood as color markers. Table A3.1. Describes these chains and the reasons for excluding them in detail. Table A3.2 provides shows the last generation output within these dysfunctional chains. Five of these chains were in the simple condition, one was in the complex condition (chains 1, 2, 10, 17, 18 and 20 in the data file).

Tab. A3.1. Descriptions of the excluded chains.

Chain ID	State of the language	Justification for removal
1	Lost one color marker and used a marker for only one color.	With one color marker missing, it became impossible to monitor the degree of overspecification.
2	Lost one color marker and used it as a clitic with all object words.	As the same marker was used throughout for both colors it could not be identified as a color marker anymore.
10	The markers were redistributed, at first chaotically and then reflecting a left-right distinction based on where the target object was located.	As the color markers changed their meanings, overspecification on the color dimension was not possible anymore and the chain could not be compared to the others anymore.
17	The markers were redistributed idiosyncratically, and were in some words left out.	Based on the signal, the learner had little chance to distinguish between colors.
18	The markers were invariably attached to specific object labels, rather than marking color.	The color markers did not distinguish between colors any more, which made color overspecification impossible.
20	One of the markers was lost and the other one is attached as a clitic to one of the object words.	The color markers did not distinguish between colors any more, which made color overspecification impossible.

Tab. A3.2. Examples of dysfunctional languages (last generation of excluded dysfunctional color chains).

			Utterance			
	Color	**Position**	**Pen**	**Cup**	**Ball**	**Book**
Starting language	yellow	left right	meeb li meeb li	fop li fop li	kur li kur li	yan li yan li
	blue	right left	meeb pu meeb pu	fop pu fop pu	kur pu kur pu	yan pu yan pu
Chain 1	yellow	left right	meeb meep	fleeb fleeb li	kur kur	yan yan
	blue	right left	meeb li meep lo	fleeb li fleeb li	kur li kur li	yan li yan li
Chain 2	yellow	left right	meep li meep li	pur li pur li	kur li kur li	yak li yak li
	blue	right left	meep li meep li	pur li pur li	kur li kur li	yak li yak li

Chain 10	yellow	left right	meeb li meeb pu	ker li ker pu	kur li kur pu	yan li yan pu
	blue	right left	meeb pu meeb li	ker pu ker li	kur pu kur li	yan pu yan li
Chain 17	yellow	left right	meeb pu meep pu	kup pu kup	kir pu kir li	yan li yan pu
	blue	right left	meep pu meeb li	kup kup pu	kir li kir pu	yan li yan pu
Chain 18	yellow	left right	meep li meep li	yon li yon li	kur pu kur pu	yon li yon li
	blue	right left	meep li meep li	yon li yon li	kur pu kur pu	yon li yon li
Chain 20	yellow	left right	meeb meeb	yan yan	kor kor	yan li yan li
	blue	right left	meeb meeb	yan yan	kor kor	yan li yan li

References

Dryer, Matthew S. 2013. Order of Adjective and Noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/87.

Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution* 1(1). 7–18.