Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language

R. Alexander Schumacher (robertschumacher2016@u.northwestern.edu)
Department of Linguistics, Northwestern University
2016 Sheridan Rd, Evanston, IL 60208

Janet B. Pierrehumbert (jbp@northwestern.edu)
Department of Linguistics, Northwestern University
2016 Sheridan Rd, Evanston, IL 60208

Patrick LaShell (patrick.lashell@canterbury.ac.nz)
NZILBB, University of Canterbury
Private Bag 4800, Christchurch, New Zealand

Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci2014) Austin, TX: Cognitive Science Society.

Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language

R. Alexander Schumacher (robertschumacher2016@u.northwestern.edu)
Department of Linguistics, Northwestern University
2016 Sheridan Rd, Evanston, IL 60208

Janet B. Pierrehumbert (jbp@northwestern.edu)
Department of Linguistics, Northwestern University
2016 Sheridan Rd, Evanston, IL 60208

Patrick LaShell (patrick.lashell@canterbury.ac.nz)
NZILBB, University of Canterbury
Private Bag 4800, Christchurch, New Zealand

Abstract

Language learners are sometimes faced with the problem of learning from input that is inconsistent or unexpected. Unexpected patterns may be typologically rare (marked) or contrary to the pattern in the first language. Using a novel game-like experimental paradigm, we examine the interaction of these factors for a set of artificial languages differing in the consistency and naturalness of number marking. The interaction of these factors in determining the degree of regularization is highly significant, and arises from individual differences that pose challenges for formal models.

Keywords: artificial language learning; grammatical number; adaptive tracking; regularization; probability matching

Introduction

Language systems are highly structured. Nevertheless, learners sometimes encounter unpredictable variation. In such circumstances, the learner must either overcome this variation, or encode it within the broader system.

Much recent work has focused on the strategies learners employ to accommodate unpredictable variation in artificial language learning. Artificial language learning is useful for seeing what learner expectations are, both for what they are learning and how structured the input should be. Two distinct strategies have been identified: learners may *probability match* or *regularize*. Probability matching occurs when learners determine the frequency of occurrence of the variants and reproduce the same variation in their output. Regularization is when learners reduce the amount of variation by favoring one variant over others. Probability matching has been observed for adults in a variety of tasks (Hudson Kam & Newport, 2005; Reali and Griffiths 2009, Vouloumanos, 2008). A number of other studies have found regularization by adult learners (Culbertson, Smolensky & Legendre, 2012; Hudson Kam & Newport, 2005; Hudson Kam & Newport, 2009; Wonnacott & Newport, 2005).

When do people regularize and when do they not regularize? Based on empirical results, Hudson Kam & Newport (2009) advance the generalization that learners are more likely to probability match variation when the number of competing variants is low and the learners are adults. Recent work by Culbertson, Smolensky & Legendre (2012) and Culbertson & Smolensky (2012) advances a different answer to this question. They propose a Bayesian model of the results of an artificial language learning experiment in which they manipulated the naturalness of word sequencing constraints (subsequently analyzed by Culbertson & Adger, 2014, in relation to the exponency of semantic scope). In their model, regularization of variable input arises if the prior expectations that participants bring to the experiment impose substantive biases. If the frequency profile of the input conflicts with these priors (as in the case of a typologically rare system), participants may shift the frequencies rather than regularizing. In this study, we investigate the interaction of frequency and naturalness in a different part of the linguistic system, namely the morphology. Though use of a novel adaptive tracking training paradigm, we also look at the time course of learning in a way that was not possible in previous studies.

The morphological contrast we examine is number marking. The system found in English, in which the singular is bare and the plural carries a suffix, is a typologically common pattern that would be associated with a strong prior in the Culbertson and Smolensky (2012) model. In an unusual pattern known as a *singulative/collective* number system, the marking is reversed. The form denoting multiple occurrences of a referent is bare and a suffix goes on the form denoting a single occurrence. Singulative number is typologically rare, but it is found in some languages, such as Welsh (Anderson, 1985). For example, the Welsh noun *adar* "birds" (the *collective* form) receives the singulative suffix *–yn* to form *aderyn* “bird". We explore the interaction of the type of number marking system (Plural vs.

Singulative) with the consistency of the input (100% vs 75% consistent). The generalization advanced by Hudson Kam and Newport (2009) leads us to expect probability matching for the 75% consistent conditions (for the 100% conditions, probability matching and regularization are not distinct). Culbertson and Smolensky (2012), in contrast, would predict that the prior bias towards the Plural system could result in regularization or shifting, depending on the input.

Our experiment uses a novel task, which is a modified *adaptive tracking* procedure. Adaptive tracking (also known as Bekesy tracking) is a common technique employed in audiology (Leek, 2001), where progress towards a threshold is determined by the responses that have been provided rather than the time course of exposure to some stimulus. The learner progresses through a series of stimuli and chooses a response after which immediate feedback is given. The learner must choose the correct response to a stimulus to proceed to the next stimulus. If the incorrect answer is chosen, the learner regresses to the previous stimulus. In our experiment, the learner is deemed to have reached threshold when at least one correct answer has been provided to every stimulus.

The adaptive tracking task was presented to the participants as a computer game, similar to many games that people play for fun. The computer game setting was selected as part of a broader research program, the Wordovators project, whose goal is to design experiments that engage participants of all ages and backgrounds. One advantage in using the modified adaptive tracking in this experiment is that it requires participants to provide a correct response to each stimulus before proceeding through the task. This enables them to build a set of accurate exemplars of the training items. Generating a guess for each item before receiving feedback also encourages the development of generalizations about the language structures. These task characteristics made it a good choice for the present experiment. In the singulative condition, the immediate feedback and focus on correct classification of every stimulus make it possible for participants to attain the training criterion, despite the expected bias towards the English plural system. The task also allows participants to quickly proceed through the task once they have learned the system that they are presented with. In many contemporary tasks, participants are required to respond to hundreds of trials. Participants able to quickly move through the task will be more engaged during the test phase than participants forced to complete boring and repetitive training. The paradigm also provides detailed information about participant performance over the whole time course of the experiment.

Using the modified adaptive tracking paradigm, we trained learners on the singulative number and English-like plural number marking systems in order to answer the question of how learners would treat inconsistency in the distinction encoded and what strategies they would employ.

Methods

Participants

Four hundred (400) participants (one hundred for each of four conditions) were recruited through Amazon Mechanical Turk over the course of two days. The large number of participants enables us to to look at individual differences in detail. All participants were native speakers of English. Each was paid three dollars.

Design

Using the adaptive tracking paradigm, participants were exposed to a miniature artificial language built around 24 image-stem pairs. Each image-stem pair had two versions. One version displayed a single token of the image, and the other displayed a group of five tokens, for a total of 48 items. On each trial, the subject saw an item together with a choice between two labels for it. Both labels had the same stem, but one label had a bare stem while the other had the stem plus an affix. Half of the image-stem pairs were used in the training phase. Half were used as novel items for generalization in the test phase. All of the training phase items also appeared in the test phase.

In the training phase, participants learned either a completely consistent marking system (the 100 condition), where the affix in the artificial language encoded the same system every time, or a 75% consistent marking system (the 75 condition), where eighteen items were one system, and the remaining six the other. The dominant marking system was either singular/plural (the Plural condition) or singulative/collective (the Singulative condition). Each participant saw each image-stem pair twice in the training phase, once with the image representing five entities and once with a single entity, for a total of 24 training stimuli.

Training stimuli were randomly assigned to groups of four to achieve block randomization over the whole experiment while counterbalancing for stimulus type, whether the suffixed form was the correct answer, and the number of entities in the images presented in that block. A fresh randomization was generated for each subject.

Phonology of the Artificial Language The words stems were five characters in length, and built using a Python script from bigram statistics drawn from the *Cronfa Electroneg o Gymraeg* ("Electronic Corpus of Welsh"). Welsh phonotactics made the words sufficiently distinct from English so as to demonstrate to the participants that they were not learning English words. The stems were paired with a suffix, which was two characters long and did not correspond to any real English suffix.

Structure of the Game Participants were given a storyline for the experiment, which was presented as a game about learning “fairy language”. Participants were told that they had to cross a river to reach the castle of the fairy "Bendith". They were told that they were going to see some words in

the fairy's language, and had to guess the correct word. The adaptive tracking procedure was visualized as planks on a bridge over a body of water to reach the castle.

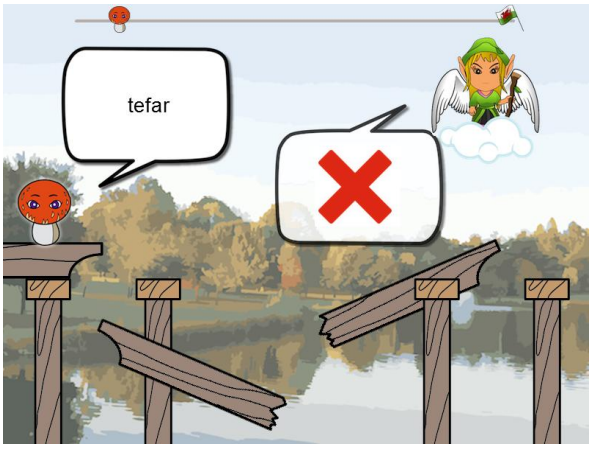

Figure 1: An example of what participants saw after picking an incorrect label on a trial.

The player advanced by providing correct answers to the stimulus presented at each length of bridge. Each correct answer was rewarded with a bridge plank, allowing the player to proceed. Incorrect answers were "punished" with the breaking of a bridge plank, and the player regressing to the previous bridge plank.

Each trial was a two-alternative forced-choice between the affixed and the unaffixed form, presented on buttons below the image. Participants clicked on the button with the word they thought was correct. Because participants were presented with both single entity and multiple entity images over the course of the game, there was a two-to-one mapping between answers that participants could provide and the marking system which those responses represented; that is, responding with the affixed form to a single entity image was considered to be a singulative/collective response, and responding with an unaffixed form to a multiple entity image was also.

Results

The measures of interest in the training phase were number of trials required to complete the training phase ("steps") and proportion correct at each training block. The measures of interest in the test phase were proportion of responses consistent with the dominant marking system (for novel test items) and proportion correct on test items that had been previously seen during training.

Training phase

Participant training performance is presented in Figure 2. Participants in the Plural 100 condition took on average 34 steps to complete the training phase, and 49 in the Plural 75. Participants in the Singulative 100 condition averaged 38 and, 53 in the Singulative 75 condition. Using a Wilcoxon rank-sum test, the difference between Plural and Singulative conditions is significant ($Z=-3.87$, $p<.0002$) and the difference between consistencies is significant ($Z=-8.21$, $p<.0001$).

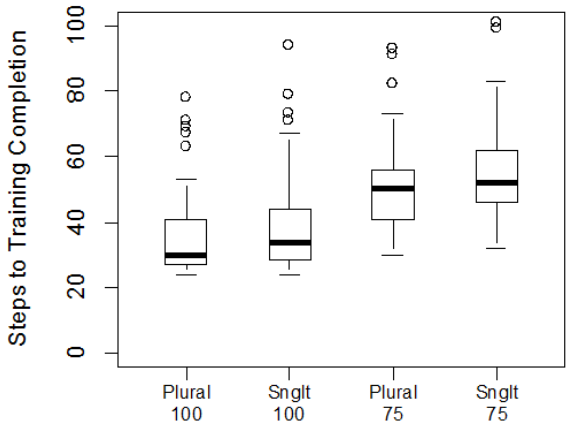

Figure 2: Boxplots of number of steps to completion of the training phase for each condition.The 100 conditions are represented on the left,
75 conditions on the right.

A player making completely random selections would take an average of 620 steps to complete the game (or a median of 479 steps), so it is evident that players performed well above chance. Absolutely perfect performance would allow the training phase to be completed in 24 steps.

Mixed logit regression was used to evaluate participant accuracy during training. These models have been found to outperform models using an arcsine transformation for the analysis of proportion data (Jaeger, 2008). They also incorporate random effects which can account for individual differences in participants and in items. Models were fit using the maximal appropriate random effects structures for both participants and item effects (Barr, Levy, Scheepers & Tily, 2013).

The fixed effects of system type, consistency, training block, and interactions were tested in the models. Block was centered to increase interpretability. Average participant performance during training is presented in Figure 3. The final model for proportion correct consisted of main effects of system, consistency, training block, and an interaction of block and consistency. Participant performance all conditions was significantly better than chance. There was a significant main effect of consistency ($b = -1.05$, $z = 14.69$, $p << .001$) showing participants in the 75 conditions were less accurate than participants in the 100 conditions. There was a significant main effect of system ($b = -0.15$, $z = 2.90$, $p < .005$), showing that participants in the Singulative condition were less accurate than participants in the Plural condition. There was a significant main effect of block ($b =$

0.47, z = 18.4, p < .001) showing that participants improved across the course of training. The interaction of system and consistency was not significant.

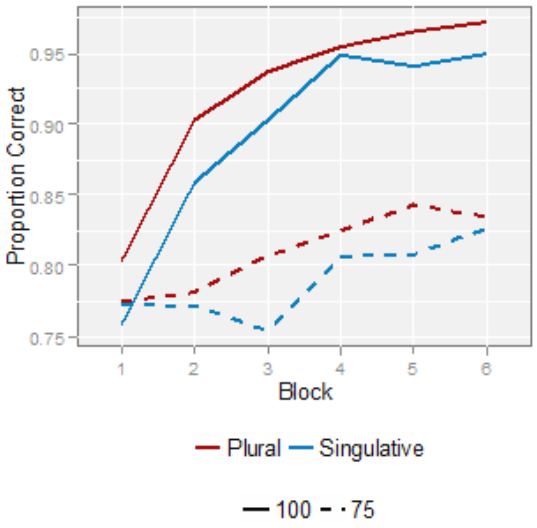

Figure 3: The average proportion of correct answers for each block during training. Performance reflects all responses at the block, including additional exposures from regressions.

In Figure 4, the time course of the training data are plotted according to the level of consistency with the dominant pattern. Consistency is identical to correctness for the Plural 100 and Singulative 100 conditions, but it is not identical for the Plural 75 and Singulative 75 conditions.

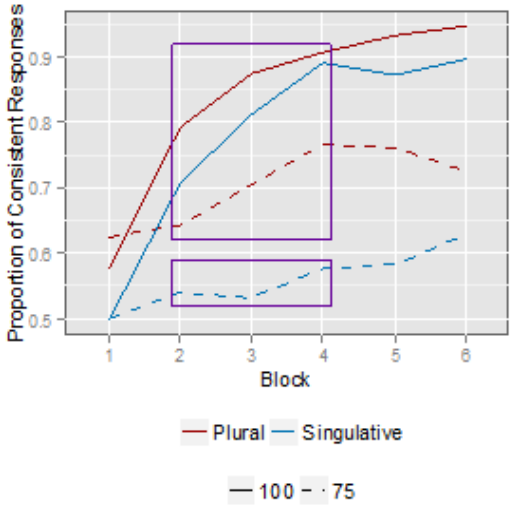

Figure 4. The proportion of consistent responses per condition per block. The highlighted area shows the blocks where the response pattern in the Singulative 75 condition is qualitatively different than that for the other conditions.

Note that the percentage of consistent responses in Block 1 was below 75% in all conditions. In Blocks 2 to 4, the response frequency climbs towards the actual frequency for three of the conditions, but not for the Singulative 75 condition where this rise is delayed.

Test phase

First, we consider average participant test performance to novel items. Results are presented in Table 1. Novel items presented in the test phase have no *correct* classification because participants had never seen them before, nor did they receive feedback. Therefore answers to novel items were scored in regards to their consistency to the marking system taught during training.

Table 1: Average Test Phase Responses Consistent with the Dominant Marking System on Novel Items

Condition	Mean	Diff. from Input
Plural 100	.94	-.06
Singulative 100	.88	-.12
Plural 75	.84	.09
Singulative 75	.54	-.21

As in the training phase analysis, mixed logit regression was used to analyze novel item performance during the test phase. The fixed effects were system and consistency, and their interaction was evaluated in the models.

The final model for the proportion of consistent responses to novel items during test contained main effects of system, consistency, and an interaction of system and consistency. There was a significant main effect of consistency, (b = -1.84, z = 5.85, p << .001) showing participants in the 75 conditions were less consistent than participants in the 100 conditions. There was a significant main effect of system, (b = -0.91, z = 2.78, p < .006), showing that participants in the Singulative conditions produced fewer responses consistent with the dominant system than participants in the Plural conditions. There was a significant interaction of system and consistency (b = -1.20, z = -2.83, p < .006) showing that participants in the Singulative 75 condition were much less consistent in their responses to novel items than predicted by the main effects of system and consistency. This interaction is evident in the low median value and large spread for the Singulative 75 condition in Figure 4A.

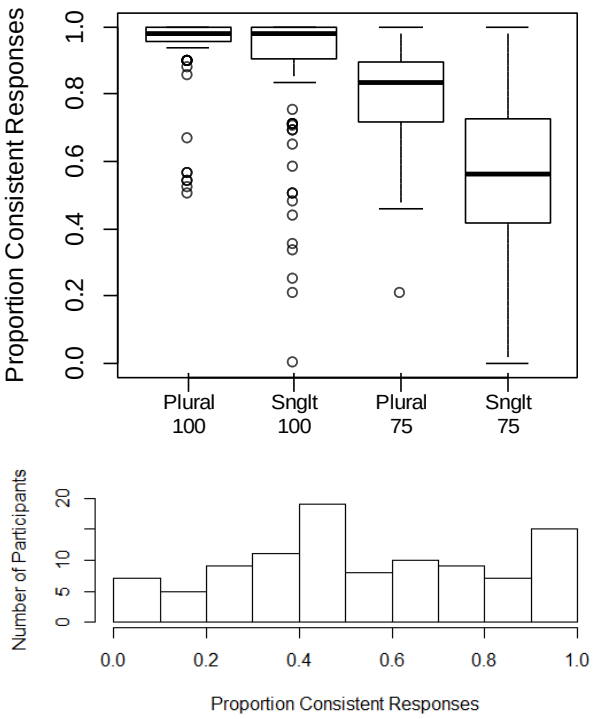

Figure 4: Upper panel: The proportion of responses on novel test items that were consistent with the dominant system. Lower panel: Histogram of individual subject response patterns in the Singulative 75 condition.

A histogram of individual response patterns in the Singulative75 condition (Lower panel in Figure 4) reveals a further pattern that is obscured in the mixed effects model and boxplots. The histogram is bimodal. One group of participants regularizes the Singulative pattern and the other shifts towards a Plural system. This is despite the fact that only two learners in the Singulative 75 provided 75% or greater consistent responses on the first training block.

A separate analysis was performed for the test items that participants had previously encountered. For participants in the 75 conditions, some of these items were from the minority marking system and were therefore inconsistent with the dominant system. An accurate response to these items in the test phase would require memory of individual words seen in training. The final model for the proportion of correct responses to previously seen items during test contained the main effects of system and consistency. The interaction of system and consistency was not significant. There was a significant main effect of consistency ($b = -2.57$, $z = 16.46$, $p << .001$) showing participants in the 75 condition were less accurate than participants in the 100 condition, and a significant main effect of system ($b = -0.69$, $z = 4.54$, $p < .002$), showing that participants in the Singulative condition were less accurate than participants in the Plural condition.

To summarize, the interaction of frequency and naturalness make the Singulative75 condition stand out from the others in several ways. For novel items there is a low percentage of consistent responses and a high level of variability, which we have traced to a bimodal distribution in the responses strategies. This behavior is not due to the memory of the training items repeated in the test phase, for which there was no interaction. Nor is it due to the initial state of the participants, as participants in all conditions began with response consistency levels lower than the levels in the training data.

Discussion

This experiment explored the interaction between inconsistency and unexpectedness in the learning of an artificial morphological system. Consistency was manipulated by contrasting a 100% consistent training condition with one that was 75% consistent. Unexpectedness was manipulated by contrasting a typologically common Plural system, which participants already know as English speakers, with an unexpected Singulative system.

During the training phase, the Singulative system proved harder to learn than the Plural system. While the Singulative 100 and Plural 100 were similar, the results for the Singulative 75 system were very different from those for the Plural 75 system. In neither condition did participants produce a probability matching pattern. This is an interesting contrast to results by Hudson Kam and Newport where adult subjects faced with inconsistency between two choices in a different task. Instead, Plural 75 participants exhibited a moderate tendency to regularize the input. The Singulative 75 participants split into two groups. One group regularized the Singulative 75 pattern, extending this pattern at rates of 75% towards 100% to novel words in the test set. The other group used the singulative/collective half the time or less on novel items in the test set. This behavior appears to reflect a strong influence of the Plural system that they had brought into the experiment from their knowledge of English. This split in the outcomes occurred only in the Singulative 75 condition.

What assumptions about individual variation might yield these results? The Bayesian model described by Culbertson & Smolensky (2012), based on Culbertson, Smolensky & Legendre (2012)'s results, is able to generate bimodal outcomes. Their model produces the bifurcation by means of prior weights, effectively previously seen trials, since the bias towards regularization in their model is constant. So, in order to produce responses that are both above and below the target frequencies, the prior must be strong enough to countermand the observations to an extent. In the case of the Singulative 75, the participants who regularized would have had to enter the experiment with a strong singulative/collective prior that would persist throughout the training. This is because the sum of the training observations would yield a probability matching effect according to their model, and only by conjunction with the

prior would the frequency of the singulative/collective be estimated at greater than input. This, however, is inconsistent with the findings from the training phase. If the learners who regularized entered the experiment with a strong singulative/collective prior, they should have produced more singulative/collective responses early in the training phase. Yet, the average proportion of consistent responses on the first block of the training for the participants who regularized the Singulative 75 was only .60. Further, they should have completed the training phase faster than participants with a Plural bias but they did not. Succinctly, if the end result according to the model is regularization of the singulative/collective, then the prior will be responsible for that regularization, but that prior would also be expected to be demonstrated throughout the training phase, contrary to fact. Interesting, this is also the case for the Plural 75, where most learners (~75%) produced more singular/plural responses than was present in the input. This group also produced less than target in the first block, at .65.

An alternate explanation for the performance on the Singulative 75 relies not on prior counts, but on how informative learners considered evidence from the different systems. On this view, all learners weight the examples they see as more or less informative, but regularizers exaggerate the majority case. Plural 75 players considered the singulative/collective items less likely to be examples of a number marking system, and so they ignored the examples of inconsistency. Conversely, although a singular/plural response was inconsistent with the dominant marking system for Singulative 75 learners, it was sporadically reinforced by 25% of the items (one item per block) and learners because of their bias consider it a likely marking system. On the novel test phase items, the players' responses could be seen as a result of what they recalled of the dominant system, combined with their propensity to respond with singular/plural marking. The disproportionate effect in the Singulative 75 of the small number of singular/plural items in supporting a persistent singular/plural bias in this condition has a suggestive link to the confirmation bias literature.

Conclusion

Participants were exposed to miniature artificial languages which represented either a singular/plural marking system or a singulative/collective marking system, with either 100% or 75% consistency.

The principle finding was of an interaction between input consistency and marking system. Participants regularized the input in the Plural 75 condition, but Singulative 75 players produced more inconsistency in their output than they were exposed to. This finding shows that the strategies which learners employ to reconcile variation depends not only on the amount of inconsistency present in the input, but on the distinction encoded by the input.

Acknowledgments

This project was made possible through a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Anderson, S. R. (1985). "Inflectional Morphology" in T.Shopen [ed.] *Language Typology and Syntactic Fieldwork* vol. III, pp. 150-201. Cambridge: Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C., &Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842-5847.

Culbertson, J., & Smolensky, P. (2012). A Bayesian Model of Biases in Artificial Language Learning: The Case of a Word-Order Universal. *Cognitive science*, *36*(8), 1468-1498.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306-329

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., &Laporte, N.(2001). Cronfa Electroneg o Gymraeg (CEG):A 1 million word lexical database and frequency count for Welsh. [On-line]

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151-195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30-66.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446.

Leek, M. R. (2001).Adaptive procedures in psychophysical research.*Perception & Psychophysics*, *63*(8), 1279-1292.

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

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning.*Cognition*, *107*(2), 729-742.

Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In *BUCLD* (Vol. 29, pp. 663-673).