People ignore token frequency when deciding how widely to generalize

Amy Perfors (amy.perfors@adelaide.edu.au)
Keith Ransom (keith.ransom@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

Many theoretical accounts of generalization suggest that with increasing data, people should tighten their generalizations. However, these accounts presume that the additional data points are all distinct. Other accounts, such as the adaptor grammar framework in linguistics (Johnson, Griffiths, & Goldwater, 2007), suggest that when the additional data points are identical, generalizations about grammaticality need not tighten appreciably: they may be made on the basis of type frequency rather than token frequency (although token frequency can affect other types of learning). We investigated what happens in this situation by presenting participants with identical data in both a linguistic and a non-linguistic context, some ten times as much as others, and asking them to generalize to novel exemplars. We find that people are insensitive to token frequencies when determining how far to generalize, though memory has a small mediating effect: generalizations tighten slightly more when people may rely on a memory aid.

Keywords: generalization; category learning; adaptor grammar; grammar learning; types; tokens; size principle; frequency

Introduction

How far should people generalize from the data they have observed? If a child points out two beagles and a basset hound and calls them GLUGGIES, should we guess that she will also call a great dane by this name? We could generalize narrowly, and guess that the word GLUGGIES refers only to small dogs. Or we could generalize broadly, and guess that it might include cats and other typical household pets.

One approach to this problem is to assume that observations are sampled randomly from the true extension of the category, and use Bayes' rule to guide inferences (Tenenbaum & Griffiths, 2001). According to this "strong sampling" scheme, if a category contains k items, each item will be observed with probability $\frac{1}{k}$. Given two hypotheses about the category that are both consistent with the observations, one narrow and the other broad, a Bayesian learner will eventually learn to prefer the narrower hypothesis as more and more data arrive. This is because the narrower hypothesis assigns higher probability to the observations. This is known as the *size principle*, and it (or something like it) has been shown to guide human generalization in a variety of contexts (Xu & Tenenbaum, 2007; Navarro & Perfors, 2010; Navarro, Dry, & Lee, 2012).

One ambiguity in this research, however, is that it assumes that all observations are independently sampled and meaningfully distinct from one another. Suppose it turned out that the "two" beagles labelled GLUGGY by the child look identical and may be the same animal. Should this be treated as one data point or two? This is a common problem: people do repeatedly encounter the same dog or hear the same sentence many times (e.g., "How are you?"). The normative treatment of identical data points depends on how one interprets the generative process behind the data. If the sampling process is something akin to drawing examples with replacement from a bag of possibilities, then the size principle should apply: seeing the same example multiple times is actually strong evidence that the true extension of the category is very small.

A more conservative approach is for people to make a distinction between distinct entities (types) and the set of instances (tokens) on which they have been observed. It is not unreasonable to assume that distinct types represent distinct samples from the category, but multiple tokens of the same type may not. The issue comes down to the *relevance* of token frequency to the inference at hand. For instance, seeing the same beagle repeatedly may be informative about the number of dogs in the neighborhood, but it does not say much about how common beagles are relative to other kinds of dog.

In linguistics this insight is captured using the *adaptor grammar* framework (Johnson et al., 2007), in which the learner makes a distinction between the underlying category to be learned (the grammar) and a mechanism that shapes the frequencies with which category members are observed (the adaptor). This framework has been successfully applied to many aspects of language (e.g., Johnson & Demuth, 2010; O'Donnell, Snedeker, Tenenbaum, & Goodman, 2011; Perfors, Tenenbaum, & Regier, 2011), and under certain parameter values it makes a different prediction than the one that emerges from strong sampling. If the same set of entities are observed many times, this is not in and of itself evidence that the true category is small: generalizations should only tighten when new *types* are observed, not new tokens.

Although the adaptor grammar approach was developed to explain linguistic phenomena, there is no reason why it should not apply more broadly. Categories can be considered to consist of a set of entities, or types (the extension of the category), and a frequency distribution can then be defined over that extension; the former is analogous to the grammar, the latter to the adaptor. On the other hand, language is different from concept learning in many ways – for instance, individual sentences are not physical entities in the same way that exemplars from a concept are. It is also possible that token frequencies are relevant in category learning problems but not language learning ones.

This paper investigates how people change their generalizations when they encounter new tokens of old types. We begin by presenting an experiment in which participants were shown a dataset of ten distinct types of exemplars, each occurring either once or ten times. Our main question is whether people tighten their generalizations when they are shown ten

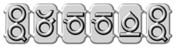
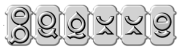
Inscription	Design
du gi bo du	
du du du bo gi la du du du	
du bo gi la la du	
bo du gi gi la bo	
pe ho sa vu vu re	

Figure 1: Sample stimuli in the INSCRIPTION and DESIGN conditions. Participants in the DESIGN condition were asked to classify bracelets with different patterns, while those in the INSCRIPTION condition classified bracelets with different inscriptions. The top two rows show training items; the bottom three show test items.

times as much data, even though the number of types is equivalent in each case. We also investigate whether domain differences exist, presenting the same data in a linguistic and non-linguistic scenario. We find that in both domains, there is little difference in generalization with increasing data. This is mediated by memory: although in all cases people generalize fairly widely, they generalize less widely when given assistance with memory.

Experiment

454 adults were recruited via Amazon Mechanical Turk. 40 participants were excluded from further analysis for failing to pass a "check" question, described below. This left 414 participants, ranging in age from 18 to 66 (mean: 31.8) and 39.4% were female. 314 of the final participants were from the United States and 68 were from India. Those remaining were from 12 other countries in Africa, North and South America, Europe, and Asia. All participants were paid $0.50US for the 5-10 minute experiment.

Procedure

The cover story informed people that as curators of a museum, they had received a collection of bracelets from their predecessor. In the first phase of the experiment, people were shown sample bracelets from the collection one-by-one, clicking Next to see the next item. The appearance and number of the bracelets, as well as whether previously-viewed ones stayed on screen, varied by condition. In the second phase people were shown new items and asked to indicate if the new bracelet belongs in the collection on a 7-point scale from "agree strongly" (1) to "disagree strongly" (7). There were 15 test items which varied according to how closely they matched the original stimuli (described in more detail below).

Conditions

This experiment varied three factors[1], resulting in a 2x2x2 design and 8 conditions. We describe each factor below.

du gi bo du
du la la gi du
du gi gi bo la du
du gi la gi bo du
du du bo du du
du du gi bo gi du du
du du la bo gi gi du du
du du du gi la du du du
du du du bo gi la du du du
du du du du bo du du du du

Table 1: Each of the 10 training stimulus types in the INSCRIPTION condition. Stimuli were generated from a grammar of the form $A^nB^mA^n$, where $A = \{du\}$ and $B = \{bo, gi, la\}$. Stimuli in the DESIGN condition corresponded exactly to these; examples are shown in Figure 1. These items occurred once each in the 1X condition and ten times each in the 10X condition.

TYPE. Two stimulus types were used, one language-like and the other non-linguistic. In the INSCRIPTION condition, participants were told that the bracelets each contained an inscription that they would read. In the DESIGN condition, participants were shown a patterned bracelet. The two stimulus types are illustrated in Figure 1. The underlying structure of the stimuli was identical in both conditions. By using bracelets as the artifacts to be learned about, the obvious linear structure in the DESIGN condition could be explained by the fact that they are bracelets, thereby minimizing the chance that people would perceive the bracelet pattern as linguistic.

QUANTITY. The major question motivating this work was whether people tighten their generalizations with additional instances of identical exemplars. We therefore varied the quantity of training stimuli people received. In the 1X condition, people saw 10 distinct stimulus types, shown in Table 1. The 'true' category is defined by a context free grammar (CFG) of the form $A^nB^mA^n$, where $A = \{du\}$ and $B = \{bo, gi, la\}$, but the 10 exemplar types are consistent with many grammars.[2]

The 10X condition differed from the 1X only in terms of the number of observations: instead of seeing each exemplar once, participants saw ten exemplars of each of the ten types. If people pay attention only to the distinct types when forming generalizations, we would expect performance to be identical in the 1X and 10X conditions, despite the fact that there is ten times more data in the latter. On the other hand, if people form generalizations on the basis of token frequency as well, we would expect them to generalize far less – to accept many fewer test stimuli as acceptable category members – in the 10X condition. Stimulus order was randomized.

MEMORY AID. Because the extent to which one generalizes is in part a function of one's memory for the training data, we varied the degree to which people had to rely on their memory to do this task. In the AIDED condition, previously encountered training stimuli were shown smaller in the

[1] We varied a fourth factor, saliency, by coloring the stimuli in some conditions. Because this manipulation did not produce effects bearing on the main point, for space reasons we do not report on it, though we include all of the data from this factor.

[2] As shown in Figure 1, in the DESIGN condition people saw patterns, not syllables. Throughout the paper, we refer to stimuli using the linguistic form (as in the INSCRIPTION condition).

Stimulus	Type
du la la gi du	Observed
du du la bo gi gi du du	Observed
du du du gi la du du du	Observed
du bo gi la la du	Depth-limited
du du du la du du du	Depth-limited
du du la gi bo du du	Depth-limited
du du du du du du bo la du du du du du du	Full CFG
du du du du du gi bo du du du du du	Full CFG
du du du du du la du du du du du	Full CFG
bo du gi gi la bo	Any order
du du du la bo du	Any order
gi du du la du la	Any order
wi sa fo	Incorrect
fo wi pe wi wi ho vu	Incorrect
pe ho sa vu vu re	Incorrect

Table 2: Test stimuli, listed in decreasing order according to how closely the match the training data. The top stimuli (Observed) precisely match stimuli that were seen in the input. The Depth-limited stimuli could have been generated by the $A^nB^mA^n$ grammar, limited to the depth of embedding as the training stimuli. The Full CFG sentences could be generated by that grammar without that limitation. The Any order stimuli could be generated by a grammar that allows A or B elements in any order; this grammar could have generated the training stimuli but also many other sentences as well. Finally, the Incorrect stimuli could have been generated by a grammar with a different underlying vocabulary.

background, and remained onscreen for the duration of the experiment. In the UNAIDED condition people saw stimuli one-by-one, with each stimulus disappearing before the next appeared. The key question of interest is whether there is an interaction between memory aid and quantity: perhaps people generalize more tightly in the 10X condition only when memory is AIDED.

Test stimuli

An essential part of this research is to be able to evaluate how tightly or loosely people generalize from the training stimuli they have seen. To that end, we constructed test stimuli that could have been generated by grammars (categories) that more or less precisely fit the input data. All stimuli are shown in Table 2, and are described in detail in this section.

Observed. These stimuli occurred in the training data. They therefore represent the tightest generalization, and we expected that participants should consistently accept them.

Depth-limited. These could have been generated by a grammar approximating the $A^nB^mA^n$ grammar, but limited to the same depth of embedding as the training stimuli.[3] It represents a tight level of generalization: people endorsing these stimuli but not full CFG would have realized that the number of elements on the left and right must match, but would not think that there could be more than four elements on either side (since that was the maximum occurring during training).

Full CFG. These stimuli could have been generated by the $A^nB^mA^n$ grammar without the limitation on depth of embedding; the left and right elements occur more often than was observed during training. As such, accepting these stimuli requires generalizing further away from the training data.

Any order. These stimuli could be generated by a grammar containing the same underlying A or B elements, but permitting them to occur in any order. Because it captures the training stimuli, it is not wrong, but accepting these stimuli amounts to generalizing quite far from the training.

Incorrect. These stimuli could be generated by a grammar with a different underlying "vocabulary" (i.e., different syllables or bracelet patterns). Accepting them requires generalizing very far from the training data. We therefore used these stimuli as a "check" to catch those participants who were not trying or did not understand the task. The 40 participants excluded from the analysis were those who agreed that these stimuli belonged in the collection (giving them a rating of 1, 2, or 3 on the 7-point scale described earlier).

Results

Figure 2 shows the average degree of generalization by each of the three main factors. Because each individual participant contributed 15 data points, standard ANOVAs were inappropriate. We therefore used three different linear mixed-effects models, one for each factor, with participant as a random effect and the factor and test stimulus as fixed effects.[4] For all three factors, there was a significant main effect of test stimulus ($\chi^2(4) = 5282.4, p < 0.0001, \eta^2 = 0.517$ for all). People responded differently to the different test stimuli, generalizing more to the ones that are more similar to the training stimuli and less to the ones that are different. This is a clear indication that they understood the task.

More relevantly to the main questions motivating this work, there is no main effect of the type or quantity of stimulus (TYPE: $\chi^2(1) = 0.002, p = 0.9576$; QUANTITY: $\chi^2(1) = 0.018, p = 0.8945$). Overall, people generalized the same regardless of whether they were classifying bracelets according to the INSCRIPTION or the DESIGN, and regardless of whether they saw ten or one hundred data points. That said, there was a significant interaction (TYPE: $\chi^2(4) = 21.49, p = 0.0002$; QUANTITY: $\chi^2(4) = 21.48, p = 0.0002$). The effect size of the interaction is extremely tiny ($\eta^2 = 0.001$ for both factors), suggesting that this effect is of negligible interest, and probably arose mainly due to our large sample size. In fact, for both factors, the model with the interaction was not preferred by BIC over the model with just the test stimulus as a fixed effect.[5] This suggests that the best model of the data is one that only includes the test stimulus, not QUANTITY or TYPE or an interaction term.

Receiving a memory aid makes a larger difference, though the effect sizes are still tiny: there is a significant main effect

[3] Because of the limitation in depth, this grammar might therefore be implementable as a regular grammar.

[4] We used the R command `lmer()` in the `lme4` library and `rsquared.glmm()` for this analysis.

[5] BICs: interaction (23162); factor and test stimulus, no interaction (23148); test stimulus only (23140); factor only (28396).

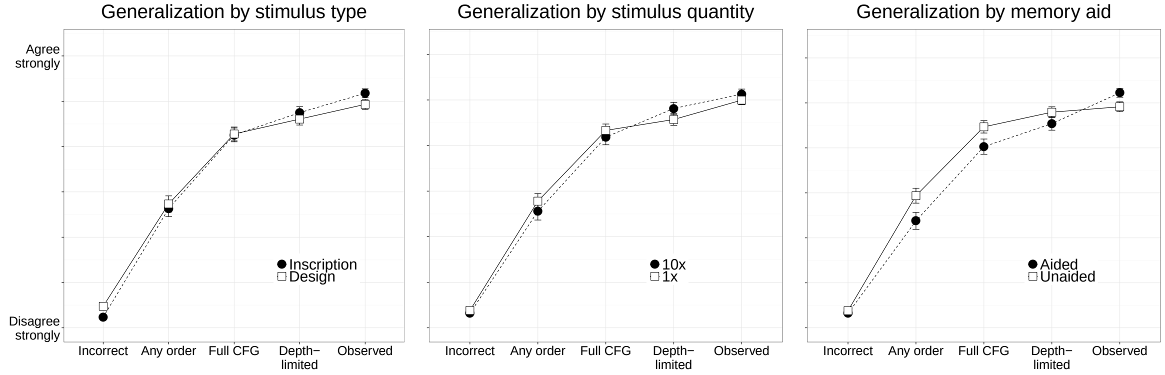


Figure 2: Generalization results by condition. The *x* axis shows the five types of stimuli, and the *y* axis shows mean responses to that stimuli. In all conditions, willingness to accept test stimuli dropped monotonically as the test stimuli grew more dissimilar. There was little variation in the willingness to generalize with changes in the type of stimulus or quantity of data. The presence of a memory aid did have an effect: people generalized more tightly when they did not have to rely on their own memory.

of having a memory aid ($\chi^2(1) = 5.32, p = 0.02, \eta^2 = 0.002$) and a significant interaction ($\chi^2(4) = 67.78, p < 0.0001, \eta^2 = 0.004$). Although the effect size remains small, the model with the interaction was preferred by BIC over any of the other models, suggesting that including the interaction this time makes sense.[6] People were more likely to accept the Observed sentences if they could see the identical training stimuli on their screen, thanks to the memory aid. They were also less likely to accept the other test sentences. This makes sense, since people who must rely on their memory may not recall if they have seen similar stimuli in training, and thus may be more willing to accept them.

Does memory mediate the effect of stimulus quantity? One might expect that people would be more affected by a greater quantity of data in the AIDED condition, because they could remember the extra data better. In fact, however, in both the AIDED and UNAIDED conditions, the results are the same: there is no effect of quantity of data (AIDED: $\chi^2(1) = 0.25, p = 0.617$; UNAIDED: $\chi^2(1) = 0.21, p = 0.650$) but a main effect of test stimulus (AIDED: $\chi^2(4) = 2476.9, p < 0.0001, \eta^2 = 0.519$; UNAIDED: $\chi^2(4) = 2872.3, p < 0.0001, \eta^2 = 0.524$). The interaction in both cases is significant, but the effect size is once again tiny (AIDED: $\chi^2(4) = 16.88, p = 0.002, \eta^2 = 0.002$; UNAIDED: $\chi^2(4) = 24.60, p < 0.0001, \eta^2 = 0.003$). Morever, BIC prefers the model with just the test stimulus over all of the others in both conditions.[7] This suggests that although having a memory aid makes people more likely to generalize more tightly overall, the effect of stimulus quantity is the negligible regardless of whether memory is aided or not.

One interesting aspect to these results is that in all conditions there is a nearly linear generalization curve for the different kinds of test stimuli. If people really were learning an underlying rule, one would expect their generalization curves to be sharper – being willing to accept all stimuli of a certain kind (e.g., all of the Full CFG stimuli) but none of the stimuli at the next level (e.g., none Any order). It is possible that the results in Figure 2, being group-level data, obscure different patterns of individual generalization.

To explore this issue, for each participant we find the grammar that best fits the response data. We do this by representing the generalization patterns of five different grammars on the test stimuli. The five possible grammars are each nested within each other: the most tightly-fitting grammar accepts only the Observed stimuli, the next most tightly fitting accepts the Observed and Depth-limited stimuli, and so forth. All grammars are named after the broadest test stimuli they accept (thus, the Depth-limited grammar fits the Depth-limited stimuli but not the Full CFG, and so on). For each participant, we calculate the fit to each grammar by taking the sum squared error between the predictions of the grammar and the person's responses. The best grammar minimizes that error.

Overall, fits to individual grammars were good. The majority of people (51.7%) had sum-squared errors of less than one (on a normalized scale). This suggests they completely misclassified only one of the 15 test stimuli or slightly misclassified a few. 81.4% had errors of less than two, and there were no errors over four.

Figure 3 shows the best-fit grammars for each condition. Strikingly, in all conditions most people are best fit by the Any order and Full CFG grammars. These are the most general grammars (aside from the Incorrect one), which suggests that people are willing to extrapolate fairly broadly from the input data. Interestingly, few people in any condition prefer the Depth-limited grammar. It suggests that as long as people perceive the dependency between the number of *du* elements, they represent this dependency at its maximum generality.

More importantly, these fits support the emerging picture that there is little effect of either quantity or type of stimulus on the generalizations individuals make. Neither factor is significant (type of stimulus: $\chi^2(3) = 5.44, p = 0.142$;

[6] BICs: interaction (23110); memory and test stimulus, no interaction (23143); test stimulus only (23140); memory only (28390).

[7] AIDED BICs: interaction (10625); quantity and test stimulus (10610); test stimulus only (10603); quantity only (13055); UNAIDED: interaction (12528); quantity and test stimulus (12520); test stimulus only (12512); quantity only (15360).

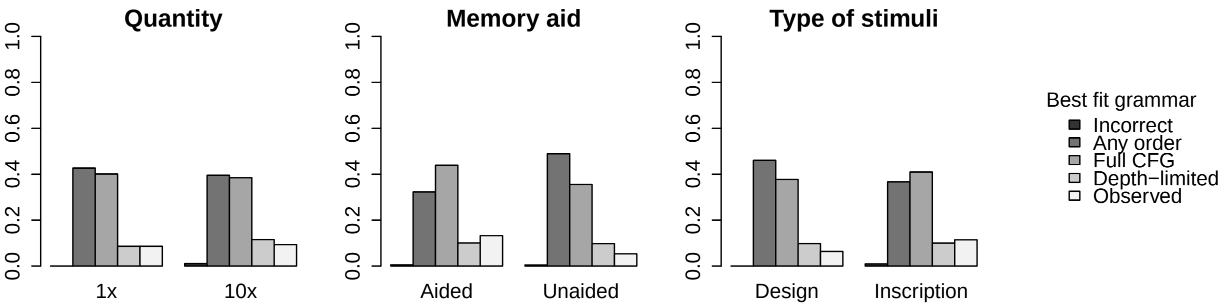


Figure 3: Fits of each individual to the grammar that best captures their pattern of responses. The y axis shows the proportion of people best fit to that type of grammar, and the legend describes the five candidate grammars. Few people in any condition were best fit by the Incorrect grammar, and most in all conditions were best fit by the two other grammars of greatest generality – Any order (which accepted stimuli in which the elements could occur in any order) and Full CFG (which is the $A^nB^mA^n$ grammar that generated the training stimuli). People's grammars do not change based on the quantity of stimuli (each type once (1X) or ten times (10X)) or the type of stimuli (INSCRIPTION or DESIGN). However, there was an effect of receiving a memory aid. When people did not rely on their own memory they generalized more tightly: fewer favored the most general Any order grammar, and more favored the tightest grammars (Depth-limited and Observed).

quantity: $\chi^2(3) = 1.23, p = 0.745$). However, as before, the existence of a memory aid does have a significant effect ($\chi^2(3) = 15.86, p = 0.001, V = 0.196$). Consistent with the previous results, people generalize less when they don't have to rely on their own memory – more people are fit by a Full CFG rather than the wider Any order grammar, which permits stimuli in which the vocabulary words can occur in any order.

As with the previous analysis, we can ask whether the effect (or, in this case, lack of effect) of quantity is mediated by memory: might there be more of an effect of seeing more data when people can remember all of the data? As before, the answer seems to be no, at least broadly speaking: as Figure 4 makes clear, the difference between the 1X and 10X conditions is not significant whether there is a memory aid or not ($\chi^2(3) = 4.00, p = 0.262$) or not ($\chi^2(3) = 2.91, p = 0.406$).

Discussion

The results in this experiment imply that people do not tighten their generalizations when they observe more data, at least not when the new observations are identical to those made previously. This lack of tightening is novel: previous work has shown that people *do* tighten their generalizations with increasing data, though sometimes less than the size principle would warrant (Xu & Tenenbaum, 2007; Frank & Tenenbaum, 2011; Navarro et al., 2012; Vong, Hendrickson, Perfors, & Navarro, 2013). However, in those studies the additional data was always new, not more instances of exemplars that had already been seen.

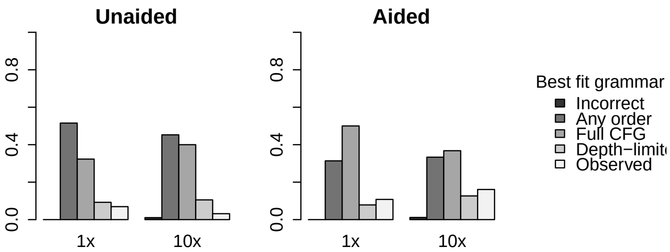


Figure 4: Fits of each individual to the best grammar, broken down by whether memory was aided. Although there is a difference between the AIDED and UNAIDED condition (as in Figure 3 above), there is no effect of quantity of data within either condition.

It is worth noting that the lack of effect of quantity or type of data in this study is unlikely to be due to a lack of statistical power; this experiment was quite large for a study in this area, with over 400 participants, and it was powerful enough that even effect sizes of $\eta^2 = 0.002$, capturing 0.2% of the total variance, were significant. By the standards of null results, this would appear to be a convincing one.

A theoretical explanation of this null result is available, as noted in the introduction: adaptor grammars. Within this formalism, the set of admissable entities (as defined by the grammar) that belong to the category is defined independently of the mechanism by which copies of previously observed are re-sampled (as defined by the adaptor). Our results are consistent with an adaptor grammar view of categories as well.

The adaptor framework also may explain why there *was* increased tightening, albeit a small amount, when people were given a memory aid. If, as is sometimes theorized, the adaptor reflects a memory cache, then removing the necessity for keeping one may result in more of an assumption that additional tokens are generated from the underlying grammar, resulting in a tightening of generalizations. Of course, as Figure 4 shows, most people still preferred the "loosest" grammars, so if this does occur the effect is not large.

The finding that people seem not to tighten their generalizations with additional token data has interesting parallels with theoretical work in the language acquisition literature (Perfors et al., 2011). This work suggests that there is sufficient evidence for a Bayesian learner given child-directed input to conclude that language has hierarchical phrase structure, but *only* if the assumptions underlying the adaptor framework are true and children tend to make grammatical inferences largely on the basis of types rather than tokens. This paper is the first experimental evidence we are aware of indicating that people are, indeed, largely unresponsive to increased token frequency when making generalizations about what other sentences are grammatical. Such behavior is also quite sensible: if people continued to tighten their generalizations as they heard more and more instances of the same sentence, one would expect adults, with decades of experience with a language, to not generalize past their input at all!

Our results show that people don't tighten their generalizations with increased quantity of identical data. However, so far we have not shown that the degree of generalization people show is matched by the degree of generalization that a normative model might prefer based only on the distribution of types. We have implemented such a model, and although space limitations prevent us from describing it in detail, it is explained in related supplementary materials.[8] Results suggest that the Full CFG grammar should be favored if learners are paying attention only to types, but the Observed grammar should be favored in the 10X condition if token frequencies are relevant. As we saw in Figure 3, most participants favored more general grammars than that. In fact, as the supplementary materials show, the majority of participants were better fit by by assuming that their inferences were type-based.

The adaptor framework applies to the domain of language, yet we found that people in the DESIGN condition, who were shown bracelets with different designs, behaved no differently than the people who saw sentences in the INSCRIPTION condition. Why might this be? One possibility is that even in the DESIGN condition, people didn't treat it as a non-linguistic task. Though we tried hard to make the stimuli look as "bracelet"-like as possible, they could still have thought of the links as a script in an unknown language. The fact that the bracelet categories were defined using a grammar-like rule may have also made the stimuli feel even more language-like. This issue was unavoidable, since we wanted a condition that was directly comparable to the INSCRIPTION condition, differing only in its surface form. However, the possibility that people treated the DESIGN condition as a linguistic one is not something we can rule out. We therefore take the lack of difference between the DESIGN and INSCRIPTION condition to be a tentative finding at this point. Nevertheless, there is no reason that an adaptor-like framework couldn't be applied to non-linguistic situations, so it is possible that people actually do behave similarly in both domains.

That said, there *does* exist some prior work in the categorization literature exploring how the frequency of identical data points affect generalization (Barsalou, Huttenlocher, & Lamberts, 1998). This work explores how people predict typicality and category membership based on how the features of an item occur with types and tokens of varying frequencies. There are two main factors that are different about our work, either of which could be the source of the difference. First, we explore *tightness* of generalization, rather than what features people attend to when forming generalizations. People might be very subtle and intelligent about when they decide whether token frequencies are relevant to the question at hand: perhaps they realize that types are relevant for determining the extent of generalization but tokens are relevant to predicting particular items (Barsalou et al., 1998) or estimating the total number of underlying types (Navarro, 2013). In terms of the adaptor framework, people might be interpreting the question in the Barsalou et al. (1998) study as being about the adaptor, in which case token frequency is the more relevant factor. The second possibility is that, as previously discussed, people may not have been treating our stimuli as being about categorization, even in the DESIGN condition. If categorization and language are fundamentally different, this could be the source of the discrepancy. Future work is necessary to tease apart these two possibilities.

Overall, however, these results indicate that people generally find increased token frequency to be irrelevant when determining how far to generalize to new examples. In particular, these generalizations tighten much less than some theoretical accounts, like the size principle, might predict. Given the vast array of evidence showing that people are sensitive to frequencies in many other kinds of situations, this is interesting. It makes sense if one assumes that learners use token frequency for some types of inferences, but generalizations about grammaticality (or, similarly, category membership) rely more on the distribution of types. This work suggests that whether frequency matters may be partially a function of what that frequency is of and what the generalization is for.

Acknowledgments

Thank you to Simon De Deyne and Natalie May for their help in designing and running pilot versions of the experiment. This research was supported by ARC grants DE120102378 and DP110104949.

References

Barsalou, L., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cog Psych.*, *36*, 203—272.

Frank, M., & Tenenbaum, J. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Johnson, M., & Demuth, K. (2010). Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Coling* (pp. 528–536).

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *NIPS 19* (pp. 641–648).

Navarro, D. J. (2013). Finding hidden types: Inductive inference in long-tailed environment. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *35th Annual Conf of the CogSci Soc* (p. 1061-1066).

Navarro, D. J., Dry, M., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(187–223).

Navarro, D. J., & Perfors, A. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*(3), 256–268.

O'Donnell, T., Snedeker, J., Tenenbaum, J., & Goodman, N. (2011). Productivity and reuse in language. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *33rd Annual Conf of the CogSci Soc* (pp. 1613–1618).

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.

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

Vong, W. K., Hendrickson, A., Perfors, A., & Navarro, D. J. (2013). The role of sampling assumptions in generalization with multiple categories. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *35th Annual Conf of the CogSci Soc* (pp. 3699–3704).

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

[8] Supplementary materials describing this analysis can be found at `http://health.adelaide.edu.au/psychology/ccs/docs/pubs/2014/perforsetal14cogsci-supp.pdf`