Sensitivity to types vs. tokens in linguistic and non-linguistic generalization

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

Insert abstract here.
Keywords: generalization; grammar learning; adaptor grammar; types; tokens; frequency; size principle

Introduction

The problem of induction, or how people generalize from limited data, is a central one in cognitive science. A core aspect of the problem is what people think about data they have not seen. For instance, if someone has seen three examples of the concept DOG (two dalmations and a terrier), they must decide if another, different item – say a spaniel – is also a dog. Similarly, people learning a language have heard certain sentences and must then decide which sentences that they haven't heard are also permitted by the grammar. The *tightness* of a generalization reflects one's willingness to extend the concept to cover increasingly different exemplars: a person willing to call a siamese cat a dog is generalizing somewhat loosely, while a person who only accepts dalmations and a terrier as dogs is generalizing very tightly. Neither person's inferences contradict any of the observed data, which implies that how tight generalizations are is guided by something other than logical necessity.

The mathematics of probability theory provide one normative standard for how generalizations should tighten with additional data. As long as one believes that exemplars are strongly sampled (i.e., sampled from the concept, as opposed to sampled from the world and then labelled), then the probability of any one data point is $\frac{1}{n}$, where n is the number of items in the concept. In other words, if there are only three possible DOGS in the world (dalmatians, terriers, and spaniel) then the probability of seeing any one of them at any one time is $\frac{1}{3}$. This principle is known as the **size principle**, and it – or something like it – has been shown to guide human generalization in a variety of contexts (CITE).

However, all of the research so far concentrates on situations in which each of the data points is distinctly different from previous ones – as if one had sampled three dogs and had been given a dalmation, a spaniel, and a terrier rather than two dalmatians and a terrier. What should people do when each data point looks superficially identical? This happens all of the time in real life: children view the same dog multiple times, language learners hear multiple utterances of the same sentence ("Hello! How are you!"), and so forth. When each data point looks superficially identical, what is the proper normative thing to do? This depends on how one interprets the generative process behind the data. If sampling is assumed to be roughly similar to drawing examples with replacement from a bag of possibilities, then the size principle should apply: seeing two dalmatians and a terrier should result in greater tightening than seeing one dalmatian and a terrier: the probability of the data in the first case would be $\frac{1}{n^3}$ while the probability in the second case is only $\frac{1}{n^2}$.

Another possibility, however, is that people might treat identical instances (tokens) as informative about the frequency distribution that they should expect to see, but only the number of different distinct types as informative about the extent to which they should generalize. This sort of inference is sensible as long as one views identical exemplars as merely multiple views of the same (single) data point. In the linguistic domain, it is captured formally in the *adaptor grammar* framework (CITE), which suggests that the learner may make a distinction between the element that licenses which items are allowed (the grammar) and the element (somewhat like a memory cache) that affects the frequency of those items (the adaptor). This framework has been successfully applied to many aspects of language learning, from morphology (CITE) to syntax (CITE). If people assume that the grammar and the adaptor are entirely separate, this implies that seeing additional identical tokens should not lead to tighter generalizations – it will affect what people expect in terms of the frequency distribution of items, but inferences about the grammar itself would only be affected by additional types. This analysis thus predicts that receiving additional identical data points should not result in tighter generalizations at all.

While the adaptor grammar framework was designed and has only been implemented in the domain of language, there is no reason that the general logic should not apply more widely. Concepts, like languages, can be decomposed into an underlying core that licenses which items are allowed in the concept (we might call this an *extension* instead of a *grammar*) which is logically separable from an additional adaptor that imposes a frequency distribution on the allowed items. That said, language is different from concept learning in many ways – most relevantly for this analysis, individual sentences are not physical entities in the same way that individual exemplars from a concepts are. Thus it may be far more natural for people to ignore the token frequency of particular sentences but not ignore the token frequency of particular exemplars of the concept DOG.

This paper investigates whether people tighten their generalizations upon seeing additional identical exemplars of previously-seen data, and whether they behave differentially if the domain is linguistic or non-linguistic. We begin by presenting an experiment in which participants were shown a dataset of 10 distinct types of exemplars, each occurring ei-

Figure 1: Sample stimuli in the INSCRIPTION and DESIGN conditions. Participants in the DESIGN condition were told that they were to classify bracelets with different patterns on them, while those in the INSCRIPTION condition were told that they were learning bracelets with different inscriptions. The goal was for those in the INSCRIPTION condition to treat the stimuli as linguistic, while those in the DESIGN condition would treat them more as objects to be classified. The top two rows show stimuli from training; the bottom three show stimuli from testing.

ther once or ten times. Our main question is whether people tighten their generalizations when they are shown ten times as much data, even though the number of types is equivalent in each case. We also investigate whether this tendency varies by domain by presenting the same situation within a linguistic and a non-linguistic (category-learning) surface form. We find that in both domains, there is no difference in generalization with increasing data. Furthermore, although generalizations tighten somewhat when people are given assistance remembering all of the items they see, a formal analysis within the adaptor grammar framework suggests that the vast majority of participants are better fit by assuming that they generalize based only on types, rather than also on token frequency. We conclude with a discussion of implications for linguistics and concept learning.

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 resulted in 414 total participants, split approximately equally between all conditions. Ages ranged 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

After completing basic demographic information, participants were told that in this experiment they were acting as curators of a museum who have some bracelets in their collection which they received from their predecessor, and which all come from the same place. All participants then answered a series of multiple-choice questions to make sure they had read and understood the instructions.

The experiment had two phases. The first was a training phase in which people were shown sample bracelets from their 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 bracelets and asked to rate on a scale of 1 to 7 whether they think the new bracelet belongs in this collection (1 = "Agree strongly" and 7 = "Disagree strongly"). There were 15 test items which varied according to how closely they matched the original stimuli (described in more detail 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 **??**. These items occurred once each in the 1X condition and ten times each in the 10X condition.

Conditions

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

TYPE. One of interest was whether people generalize differently depending on whether they see the stimuli as linguistic or not. In the INSCRIPTION condition, participants were told that the bracelets each contained an inscription that they would read. In the DESIGN condition, participants saw the jewelled pattern of the bracelet, which was designed to be of similar size and complexity as the inscription (see Figure **??**). In order to make the conditions comparable the patterns in the DESIGN condition directly corresponded to the inscriptions in the INSCRIPTION condition. However, we hoped that the cover story and bracelet-like appearance of the stimuli would make this as little like a linguistic task as possible.

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 **??**. Each type was generated from the grammar $A^nB^mA^n$, where $A = \{du\}$ and $B = \{bo, gi, la\}$.[2] This grammar is a context-free grammar (CFG), but as we will see in more detail, the specific 10 stimulus types can be captured with varying degrees of exactness by grammars of various sorts; which grammar is inferred will affect which additional stimuli are accepted as members of the same category. As is typical with a CFG, there are fewer stimuli at higher depths of embedding.

The 10X condition was identical to the 1X condition except that people saw 10 exemplars of each of the 10 types

[1] We varied a fourth factor, saliency, by varying whether they were colored or not. Because this manipulation did not produce interesting effects, for space reasons we do not report on it.

[2] In the DESIGN condition people saw patterns that exactly corresponded to these syllables, as in Figure **??**. For ease of reference henceforth we will just refer to them as they appear 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.

(100 stimuli in total). If people are paying 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. The order of all stimuli was completely random.

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 MEMORY-AIDED condition, each of the training stimuli remained onscreen after the participants clicked Next; previous stimuli were shown smaller (but still legibly) in the background. They remained onscreen while the test questions were being answered as well. The MEMORY-UNAIDED condition was more like a typical category-learning experiment: people saw each stimulus one-by-one, and it disappeared before the next stimulus appeared. Of interest is whether people generalize less in the MEMORY-AIDED condition, in particular if there is an interaction with the QUANTITY or TYPE manipulations.

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 that more or less precisely fit the input data. All stimuli are shown in Table **??**, 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 (since there were never more than four during training).

Full CFG. These stimuli could have been generated by the $A^nB^mA^n$ grammar without that 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 with the same underlying A or B elements, but permits 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 **??** shows the average degree of generalization by each of the three main factors. Two-way ANOVAs showed that for all three factors, there was a significant main effect of test stimulus (TYPE: $F(4,6200) = 1662.6, p < 0.0001$; QUANTITY: $F(4,6200) = 1662.6, p < 0.0001$; MEMORY: $F(4,6200) = 1678.3, p < 0.0001$). People responded significantly 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 as we expected and is a clear indication that people 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: $F(1,6200) = 0.01, p = 0.913$; QUANTITY: $F(1,6200) = 0.07, p = 0.786$). 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 10 or 100 data points. That said, the interaction for both factors was significant (TYPE: $F(4,6200) = 4.14, p = 0.002$; QUANTITY: $F(4,6200) =$

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

Figure 2: Results.

4.13, $p = 0.002$). Based on Figure **??**, it appears that people are slightly more willing to accept the Depth-limited stimuli in the INSCRIPTION and 10X condition, but slightly less willing to generalize more broadly in those conditions. The effect size is tiny, suggesting that these differences are small.

Receiving a memory aid, however, makes a larger difference: there is a significant main effect of having a memory aid ($F(1,6200) = 22.52, p < 0.0001$) and a significant interaction ($F(4,6200) = 13.12, p < 0.0001$). Not surprisingly, 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 is also not surprising, since people who must rely on their memory may not recall clearly if they have seen stimuli like the test stimuli in the 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 MEMORY-AIDED condition, because they could remember all of the extra data better. Within the MEMORY-AIDED condition, there is a main effect of test stimulus ($F(4,2825) = 766.0, p < 0.0001$), no effect of quantity of data ($F(1,2825) = 1.1, p = 0.286$), and a small interaction ($F(4,2825) = 3.1, p = 0.014$). Within the MEMORY-UNAIDED condition, the main effects are the same (test stimulus: $F(4,3365) = 932.6, p < 0.0001$; quantity: $F(1,3365) = 0.8, p = 0.369$) but the interaction is stronger ($F(4,3365) = 4.9, p = 0.0006$). This suggests that although the basic effect of stimulus quantity holds regardless of whether there is a memory aid, it is exacerbated when there is not.

So far these findings suggest that people generalize differently when given additional identical data or when the stimuli look different.

Now motivate the modelling analysis, as a desire to see if more people are more type-based or token-based? Look at other interactions? (Will take up a lot of space).

Discussion

Was the bracelet in the *design* condition really non-linguistic, or did people interpret it as a script in an unknown language?

Something about the adaptor explaining the memory aid stuff

Barsalou et al appeared to find that generalisations about categories were on the basis of tokens. they were asked about specific items (one whose features were on the higher-frequency individual token, one whose features matched more of the types). the questions were which one is more likely to belong in the category and which one is more typical. in general it feels like they are doing something different here - what? aren't looking at tightness of generalisations. tightness of generalisation is about in general how similar the items have to be before they are okay - what we show is that token frequency doesn't affect it that much. they are looking at whether specific items, which match more on tokens or types, are likely to be included. if people are interpreting that question to be about the adaptor part then it is totally possible to observe that. also, it's a very different kind of category (and possibly not entirely category-like).

Also should draw the link to my stuff on hierarchical phrase structure. Point out it only makes sense, because generalising by tokens implies that as people get older they should be increasingly unwilling to generalize at all beyond what they have heard (token frequency gets immense). But we know people of all ages are willing to generalize new constructions.

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.