Published in final edited form as:

J Mem Lang. 2010 February 1; 62(2): 98–112. doi:10.1016/j.jml.2009.10.002.

Category induction via distributional analysis: Evidence from a serial reaction time task

Ruskin H. Hunt and Richard N. Aslin

University of Rochester

Abstract

Category formation lies at the heart of a number of higher-order behaviors, including language. We assessed the ability of human adults to learn, from distributional information alone, categories embedded in a sequence of input stimuli using a serial reaction time task. Artificial grammars generated corpora of input strings containing a predetermined and constrained set of sequential statistics. After training, learners were presented with novel input strings, some of which contained violations of the category membership defined by distributional context. Category induction was assessed by comparing performance on novel and familiar strings. Results indicate that learners develop increasing sensitivity to the category structure present in the input, and become sensitive to fine-grained differences in the pre- and post-element contexts that define category membership. Results suggest that distributional analysis plays a significant role in the development of visuomotor categories, and may play a similar role in the induction of linguistic form-class categories.

Keywords

statistical learning; category induction; serial reaction time; distributional analysis; artificial grammar

Nearly all temporally ordered behaviors involve some form of prediction (Lashley, 1951). This is true for behaviors ranging from simple conditioned reflexes to complex cognitive and motor skills. Indeed, one aspect of successful, adaptive behavior is the ability to reduce uncertainty about upcoming events. In temporally ordered behaviors, prediction is based on previous experience and serves to anticipate the next event in a sequence, thus facilitating the preparation of an appropriate response. Prediction can be based on two fundamental properties of the input: the temporal co-occurrences of elements (surface statistics) and the likelihood that one set of elements is followed by another (category statistics). For a small corpus, the relevant surface statistics are tractable, but as the number of elements grows, it becomes increasingly improbable that a learner can keep track of these statistics and utilize them predictively in real-time. Thus, the ability to form categories and make predictions at this higher level reduces the computational load on the learner.

The potential utility of a predictive mechanism for learning the types of structures found in natural languages is the goal of the present experiments. We are interested in assessing not just

© 2009 Elsevier Inc. All rights reserved.

Corresponding author: Ruskin H. Hunt, Institute of Child Development, 51 East River Road, University of Minnesota, Minneapolis, MN 55455, hunt011@umn.edu, Vox: 612-396-5733, Fax: 612-624-6373.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

whether, but how distributional information is used to form the particular type of category known as a linguistic form class. Unlike many other types of categories, the relationship between form class membership and the perceptual features of category members is arbitrary, and for many form classes, membership is independent of the surface perceptual features of category members. Instead, membership in a category, and perhaps the formation of the category itself, is thought to be, at least in part, a function of the distributional evidence that accumulates over time in a given linguistic setting. However, natural languages contain many sources of distributional information (e.g., phonetic, phonological, lexical, semantic, pragmatic) that are virtually impossible to balance in an experimental task. Thus, we employ a non-language visuomotor task that allows us to control precisely the distributional properties of the input. Reaction times to specific stimuli in familiar and novel sequences allow us to assess learning of the surface statistics of the input, as well as generalization from the training set based on acquisition of the underlying rules and formation of visuomotor categories.

Form-class Acquisition

Form-class (or grammatical) categories are the basic units out of which most grammars are built. Hence, their formation is central to the question of how learners acquire language. Most grammars, and phrase structure grammars in particular, consist of rules for ordering or manipulating abstract categories of elements, such as nouns, verbs, and so on (Chomsky, 1955/1975, 1957). The elements that may appear in those categories are individual words. One consequence of form-class category membership is that it licenses the use of a given word in other cases where that category occurs (Radford, 1988). Such grammatical, within-category substitutions contrast with ungrammatical, cross-category substitutions. Crucially, the rules for constructing sentences in this type of grammar involve manipulating categories, not specific elements. Thus, in order to utilize a phrase structure grammar, one must have access to the categories that are used in the grammar as well as the words that can occur in those categories. For a language learner, this must entail the ability to infer which words belong to the different categories in the grammar, and possibly the category types required by the grammar as well.

Evidence for the existence of a form-class category may be inferred based on how a learner responds to four types of novel conditions. In the first case, learners should reject novel exemplars in which an element from a different category occurs in a target category context (e.g., the use of a verb where a noun should occur). In the second case, learners should generalize to novel exemplars in which the individual elements in the exemplar are familiar, but the specific combination of elements is novel. In the third case, learners should generalize an exemplar encountered in one surrounding context that licenses a particular form-class category to another surrounding context that also licenses that form-class category (e.g., the use of a noun that was previously encountered only as the object of a preposition to another context that also licenses nouns, such as the complement of a determiner). In the fourth case, learners should generalize exposure to a novel exemplar in a particular form-class context to a different form-class context for which that form-class is licensed (e.g., after encountering the nonce “dax” as the object of a preposition, the use of “dax” should generalize to use as the complement of a determiner). In the present experiments, we use the first two criteria to assess the likelihood that learners induce categories based on the distributional information available over a corpus of input.

Artificial Grammar Learning

A related literature uses artificial grammar learning (AGL) paradigms (Reber, 1993) to investigate the learning of probabilistic sequences. Reber (1967) presented participants with strings of visual stimuli (i.e., letters) organized according to a finite-state grammar.¹ Each string was generated by a single pass through the grammar and was presented simultaneously,

as a single printed item, rather than sequentially. Reber demonstrated that participants who were exposed to strings generated by such a grammar could discriminate between grammatical and non-grammatical test strings, even when both types of test strings were novel.

In a recent review Pothos (2007) summarized over 120 published AGL studies and nine different models that have been proposed to account for the diverse empirical results. Much of this literature focuses on the ability of learners to discriminate between novel strings that conform to an underlying finite-state grammar from those that violate it. Importantly, a finite-state grammar is typically construed in terms of surface elements, not categories with context-dependent licensing. Thus, when a learner correctly judges a previously unattested transition as grammatical, that induction must be based on a “local” inference (e.g., A-B-B-C and A-B-D are legal, therefore A-B-B-D must be legal as well). However, inferring the existence of a pattern, or even a set of rules, over a corpus of input, is different from inducing categories that exhibit full licensing for novel category members. This is a key distinction between finite state grammars and phrase-structure grammars. Studies of AGL in children presented with a phrase-structure grammar suggest that surface cues correlated with phrase boundaries (e.g., pitch) enhance acquisition (Braine, 1963; Morgan, Meier & Newport, 1987), although learning is also possible with distributional information alone (Saffran, 2001; Reeder, Newport & Aslin, 2009).

Reber (1969) found that changing the specific elements (surface structure) used in finite-state grammar test strings produced little or no interference in a transfer task, whereas changing the rules of the grammar did (though compared with non-transfer conditions, most transfer experiments do show a decrement in grammaticality accuracy, e.g., Brooks and Vokey, 1991; Gomez and Schvaneveldt, 1994; Dienes and Altmann, 1997). He argued that what participants had become sensitive to was not the order of specific sequences of surface elements, but rather their underlying grammatical structure. However, the only way that a learner could induce the element-substitution rule is by relying on element repetitions, such that A-B-B-C maps onto E-F-F-G (see Gomez, Gerken & Schvaneveldt, 2000). A similar ability to induce the underlying rules of a simple pattern of elements (A-A-B or A-B-A) when the elements are novel has been shown for 9-month-olds infants by Marcus, Vijayan, Bandi Rao, and Vishton (1999).

More relevant to the current set of experiments is the work of Gomez and Gerken (1999) who used element strings with optional repetition. They showed that 12-month-old infants discriminate between novel grammatical strings and novel ungrammatical strings. But again, this was not a study of category learning because the finite-state grammar was not organized into form-classes. In a verbal AGL study with adults, Monaghan, Chater, and Christiansen (2005, Exp. 4) manipulated frequency of occurrence (as a proxy for all distributional information) and phonological similarity for the elements of categories following category markers. They found that high frequency words show better evidence of category membership than low frequency words in the absence of phonological cues. Mintz (2002) explored the ability of learners to form categories using an auditory nonsense-language in which distributional cues were the only information source available for category induction. Participants were given recognition and confidence-rating tests using 1) previously encountered strings, 2) strings withheld from the exposure session, 3) control strings containing a familiar word in a familiar location, but in a novel (though familiar) surrounding context, 4) and strings with words in locations that violated the category structure of the grammar. Learners

¹A finite-state grammar is a method for representing the space of possible strings that can be generated from a particular set of rules. It consists of a set of nodes and links between those nodes. The links define the possible transitions that may occur between nodes, and each link has an associated probability of occurrence. Strings are generated from the grammar by beginning at a starting node and then progressing along links to other nodes according to the probabilities of the links.

showed differential performance between withheld and ungrammatical strings, as well as between withheld and control strings. Importantly, no differences were found between withheld strings and grammatical strings attested in the input. Better recognition of withheld strings, as compared to control strings, could not have occurred based on absolute sentential word position alone, because this was equated for these two string types. The conclusion was that performance in the case of the withheld strings must have occurred based on generalization of word category membership. For this to occur, learners must have engaged in distributional analysis of the input corpus.

However, these studies did not quantitatively assess or manipulate the specific statistics over which distributional analysis occurred, leaving open the question of what aspect of the available distributional information learners utilized in service of category formation. In natural language acquisition, learners encounter only a small number of the possible word combinations that could be evidence for the presence of a form-class category. Thus, learners are faced with the problem of inferring category membership based on a limited amount of sparse data. There are a number of statistical cues that learners could utilize toward this end, including item repetition, positional order, and co-occurrence. Some sources of information in a large corpus of input, such as immediate repetition or serial positional order of words, may provide highly salient cues to form-class membership. However, it is unlikely that these cues alone can fully characterize the complexity of form-class categories in natural languages, or that they suffice to permit learners to induce those categories based on a limited set of input. A more likely candidate for accumulating evidence of form-class structure across a corpus of highly variable input would be the transitional statistics among words. If learners possessed a computational mechanism capable of tracking the co-occurrence statistics of individual words, they might be capable of inferring the co-occurrence contexts in which certain classes of words may occur.

A number of researchers (Kiss, 1973; Redington, Chater, & Finch, 1998; Mintz, 1996, 2002, 2003; Mintz, Newport, & Bever, 2002) addressed the question of whether there is sufficient distributional information in child-directed speech for a learning mechanism to induce form-class categories similar to those found in English by using distributional analyses of transcribed corpora of child-directed speech together with hierarchical clustering analysis techniques. In each of these cases, the distributional analysis of the input corpus keeps track of the co-occurrence patterns of target words and potential context words, thereby generating a record of the number and kind of contextual co-occurrences that appear for each of the target words in the input corpus. Depending on the number of context words before and after a target word that define the context, the size of the set of target words and context words, and whether only the most frequent target words are considered, these computational models achieve reasonable success in selecting clusters that correspond well to syntactic form-class categories.

It remains to be determined, however, what kinds of co-occurrence statistics might *minimally* suffice for human learners to show evidence of form-class induction. Although a large (perhaps infinite) number of computational statistics might be employed to analyze a given set of input, perhaps the simplest statistics that track co-occurrence are transitional statistics spanning two immediately adjacent words, or elements. If category induction could be demonstrated based on co-occurrence contexts for which the only evidence was tightly controlled pair-wise statistics, it would indicate the existence of a basic computational mechanism operating in the service of an abstraction that generalizes beyond the specific exemplars used in its formation. If evidence of form-class induction can be found in such a case, then the specific statistics that have been characterized may be manipulated to assess their relative influence on category formation. Thus, the specific characterization of pair-wise transitional statistics and an assessment of whether and how they serve the purpose of form-class category induction is the aim of the present work.

The Serial Reaction Time Task

The task employed in the present experiments involves a multi-choice, disjunctive, serial reaction time (SRT) paradigm (see Nissen and Bullemer, 1987). In this paradigm, a visual cue is linked to a distinct, spatially-specific, motor response. Typically, the participant presses a button that corresponds to one of 4–5 visual cues whenever the cue is illuminated on a computer screen. If the sequence of cues is random, then participants show an overall improvement in reaction time (RT) over trials. However, if sequential structure is embedded in the stream of cues, then RTs decline below this baseline level. Thus, the dependent measure of learning is the pattern of RT differences that correlate with the underlying temporal-order statistics.

SRT experiments originally involved relatively few cues whose temporal order was fixed and repeated without interruption (e.g., Cohen, Ivry, & Keele, 1990), thereby creating a perfectly predictable, deterministic sequence. With continued practice, learners show faster RTs to the sequence than to brief, untrained modifications to the sequence. Cleeremans and McClelland (1991) created probabilistic sequences analogous to the AGL paradigm. A finite-state grammar generated transitions from one cue to the next, with occasional random substitutions of a grammatical element to create an ungrammatical transition. This prevented participants from perfectly anticipating the cues and acquiring explicit knowledge of the sequence. Nevertheless, responses to grammatical transitions had faster RTs than did responses to ungrammatical transitions, indicating that implicit knowledge of the grammar had been acquired. Although many researchers have varied the level of statistical information in a sequential stimulus stream or the reliability of that information (e.g., Cohen, Ivry, & Keele, 1990; Cleeremans & McClelland, 1991; Stadler, 1992), they typically rely on aggregate differences in predictability (i.e., pooled RTs, regardless of differences in predictability among transitions) to demonstrate learning. This confounds a number of different sources of statistical information. Thus, most applications of the SRT paradigm, whether using deterministic or probabilistic sequences, have not explored the specific information participants use to learn the underlying temporal structure.

Hunt and Aslin (2001) adapted the SRT task to address which statistics learners can extract from input sequences. In their task, pairs or triplets of elements (analogous to syllables in words; see Saffran, Newport, and Aslin, 1996) were randomly concatenated into a continuous sequence. RTs were used to assess participants' learning of element predictability. The statistical structure of these sequences differed from previous SRT studies in that there was neither a single deterministic sequence of elements nor a probabilistic sequence that was compared only to a random sequence. Both low and high predictability element sequences occurred in the input. RTs to low predictability elements were reliably slower than to high predictability elements. They also determined that participants can exploit more than one statistic in order to learn the predictiveness of elements. In contrast to this paradigm, in which the sequence of visual cues was continuous, the experiments reported below presented short sequences separated by pauses, analogous to multi-word utterances in natural languages.

Overview of Experiments

The experiments utilized the SRT paradigm to mimic certain aspects of the structure of natural language input, even though the specific response required of the learner was quite different from what occurs in natural language learning. Because the SRT paradigm allows one to control the statistics of the input to which the learner is exposed, we could assess whether learners are capable of exploiting the statistics of the input to infer category membership. Two experiments were conducted, each of which involved responding to sequences of visual stimuli generated by small, artificial grammars. The grammars and the stimuli were designed to have no information content other than the distributional properties of the stimuli within sequences

(i.e., no phrasal, semantic, prosodic, phonological, referential, or other cues). All stimuli consisted of unique single shapes that were presented in strings of temporally ordered sequences.

The experiments were designed to present learners with a set of training strings containing most, but not all, of the possible sequences that could be generated from a given grammar. After training, learners were presented with additional strings, some of which contained transitions that were novel, but which did not contain elements that violate putative category membership (generalization strings). Other strings also contained transitions which were novel, but contained elements that *did* violate the putative category membership experienced during training (illegal category extension strings). The question was how learners would react to these novel strings. If they had begun to induce contextual categories, they could only have done so on the basis of the distributional information in the training input. Evidence of having induced categories would be a tendency for participants to treat novel strings with legal extensions in a manner similar to grammatical training strings, and to treat novel strings with illegal extensions as different from strings encountered during training.

The first experiment tested the feasibility of the experimental paradigm and established a baseline for the RT measure of learning. A variety of statistics that characterize the transitions of elements in the sequences was carefully controlled so that category induction could be assessed. The second experiment addressed the issue of context density by increasing the number of transitions that define a category. “Context density” refers to the number of unique transitions that define the distributional context of a category. In the current experiments, a target category (X) is defined by the distributional pattern its elements share with elements from the preceding and following categories. In Experiment 1, only a single transition occurred between any given element in the category preceding category X and the elements within category X. Experiment 2 manipulated the grammar to eliminate such unique transitions. However, doing so increased the number of elements in the preceding category, which meant that the number of possible transitions into category X also increased. From the perspective of a learner, therefore, the evidence for category X would be distributed over a wider number of transition types. Although this presumably means that a learner receives less evidence of category structure over any given amount of input, it may still result in robust category formation. This is because the learner is less likely to notice any single transition type if there are a large number of types that define the category, and as a result are less likely to interpret any missing transition as a true “gap” in the grammar. On the other hand, if the transitional statistics become sufficiently sparse given high context density but few examples of each context, then generalization should become weaker as gaps are judged to be “real”.

Experiment 1

The goal of Experiment 1 was to determine whether learners show evidence of having induced categories from an input corpus based on distributional information alone, and to characterize the nature of that information. Of course, a variety of statistics may be extracted from an appropriately structured sequence of stimuli (e.g., Reed & Johnson, 1994), including surface statistics (e.g., the predictability of element transitions; Cleeremans & McClelland, 1991) and category information (e.g., a set of abstract rules; Reber, 1989). Even for the simple case of two successive elements in a temporal sequence, there are a number of different statistics that characterize the relation between those elements. One simple statistic that could be used to generate predictions about element occurrence is overall element frequency. A related statistic is element probability (i.e., $P(X)$), which normalizes the frequency of each element by the total frequency of elements in the sample. Human learners can readily keep track of either of these statistics (Hasher & Zacks, 1984). Another statistic is bigram frequency, and the related statistic bigram relative frequency (RF; also known as bigram probability, $P(XY)$), which normalizes

bigram frequency by the total frequency of bigrams in the sample. Both bigram frequency and bigram relative frequency imply the extraction of temporal-order information, and one or both of these statistics has been shown to underlie the behavior of adults (Saffran, Newport & Aslin, 1996) and infants (Saffran, Aslin & Newport, 1996) in a word-segmentation task.

One potential problem with bigram frequency or bigram relative frequency in learning temporal-order information is that across a corpus with many transitions between elements, some transitions may occur equally often but vary widely in how many other alternatives follow a given element. A statistic that better reflects the predictability of element co-occurrence is forward conditional probability (FCP; i.e., $P(Y|X)$), which normalizes the RF of a pair of elements by the probability of the first element in the pair. Evidence for such extraction has been provided in the auditory (Aslin, Saffran & Newport, 1998) and visual (Fiser & Aslin, 2002) domains. Furthermore, and consistent with work in the animal learning literature (e.g., Rescorla & Wagner, 1972), Hunt and Aslin (2001) showed that participants in a SRT task can use a conditionalized statistic (i.e., FCP) as the basis for estimating predictability. There are many situations in which frequency-based statistics can lead to errors in estimating predictability (e.g., when the frequency of some elements is very low). Under these circumstances, FCPs provide a more reliable estimate of predictability than either frequency or RF. A less intuitive conditionalized statistic is backward conditional probability (BCP; i.e., $P(X|Y)$), which is the probability that the previous element was X given that the current element is Y. Whereas FCPs are relevant in the distributional analysis of both segmented input and input streams with no pauses, BCPs may be especially useful in cases where the input occurs in shorter, segmented units, such as the strings used here. Although spoken language utterances may, at times, be quite lengthy, they are never continuous (i.e., without pauses). Continuous input provides no obvious structural position from which to begin a backwards computation. It may be that pauses between strings provide learners with a constrained unit over which to perform relevant computations, which may render BCPs more useful than in the case of a continuous, unsegmented input stream.

Method

Participants—Data were collected from 20 adults (13 men and 7 women; 17 right handed and 3 left handed) ages 18 to 29, all of whom were undergraduate or graduate students at the University of Rochester. Participants were paid \$90.00 for approximately 4.5 hours of participation over the course of 5 days, regardless of performance.

Apparatus—Learning was assessed using a touchscreen (17-inch LCD, 75Hz refresh, 1024×768 resolution) that was mounted on a customized base, putting the screen at lap level and at a 45-degree angle relative to the ground so that it was comfortably within arm's reach. Custom hardware interfaced the touchscreen with a microcomputer. Custom software controlled the temporal and spatial sampling of the touchscreen and could detect touches with sub-millisecond accuracy. Spatial localization accuracy was ± 0.2 inches.

The visual display consisted of two parts (see Figure 1A). One was a response frame of 16 shapes. Each shape was located in the center of a 1.5-inch-square box, with .375 inches between boxes. For right-handed participants the frame was offset to the right; for left-handed participants it was offset to the left. The other part of the display was a stimulus box in which individual shapes appeared. This stimulus box was offset in the direction opposite the frame.

Design—The design of Experiment 1 was based on an artificial grammar composed of 6 categories (see Figure 2A). Each category was comprised of 2 or 5 unique elements. For any given participant, each number in Figure 2 corresponded to a unique shape and location in the response frame of the visual display. The specific shapes and their locations in the response

frame were randomly selected for each participant. There were only 15 elements in this design despite the fact that there were 16 elements in the response frame. Although the sixteenth shape was visible in the response frame throughout the experiment, it never occurred as a stimulus item. Stimulus strings were generated by progressing through the grammar from left to right, choosing elements from the appropriate column at each step, and displaying them one at a time in the stimulus box (Figure 1B). Category B was optional and occurred 50 percent of the time. This optionality was included so that the serial order position of the elements that were legal and illegal members of category X would vary. It also resulted in a larger number of string types.

Three string sets were created in Experiment 1. The training set consisted of strings generated by the grammar, with the exception of element transitions indicated by the dashed lines in Figure 2A. To ensure that no participant encountered the full set of training strings that could be generated by the grammar, a subset was removed and never presented. The strings that were removed were selected so as not to alter the statistics of the training set. The generalization set contained strings generated by the grammar using element transitions indicated by the dashed lines in Figure 2A. The illegal category extension set consisted of strings that violated the category structure of the grammar. These strings were produced by substituting the elements normally found in category X with the elements found in category E (with the constraint that the element that occurred in category X could not be repeated in category E; see Figure 2B). Illegal category extension strings and generalization strings were used as test strings on Day 5 to assess category learning and generalization, respectively.

The grammar in Figure 2A is capable of generating 240 unique string types.² Of these, there are 192 possible training strings and 48 untrained generalization strings. Of the 192 possible training strings, 156 form the training set, balancing a variety of crucial statistics. The illegal extension set contains 48 string types. Each set contains a 2:1 ratio of string types that contain category B to those that do not. In order to produce sets with equal numbers of tokens that do and do not contain category B, the non-category-B string types were doubled in each set. Thus, the training set contained 156 string types and 208 tokens, the legal extension set contained 48 types and 64 tokens, and the illegal extension set contained 48 types and 64 tokens.

The constraints used to generate the training and test strings produced four subsets of transitions between elements in category C and elements in category X (for the purposes of calculating statistics, category X was divided into four corresponding subtypes, X1, X2, X3, and X4). Two subsets of transitions were encountered during training, and differed only in backward conditional probability. These low and high BCP training transitions (found in training strings) were comprised of familiar category-elements in familiar element-transitions that were attested in the input on Days 1–4. For low training transitions in Experiment 1, a given element from category X could be preceded by more than one element of category C. In contrast, high training transitions in Experiment 1 consisted of unique transitions in which an element from category X was preceded by only one element of category C. From the perspective of a learner, it is not clear if high training transitions should generalize to other, unattested transitions, or whether they represent gaps (i.e., exceptions) in the grammar. Generalization transitions (found in generalization strings) contained familiar category-elements in unfamiliar element-transitions that never occurred in the training input. These were the key transitions that tested the ability to generalize element category membership. Illegal transitions (found in illegal category extension strings) were composed of unfamiliar category-elements in unfamiliar element-transitions. These transitions were clearly ungrammatical. All of the category and category transition subtypes (both training and testing) along with their constituent elements and element transitions are given in the left half of Table 1.

²A string type is a particular sequence of elements (shapes). In contrast, a string token is one instance of a given string type.

The presence of categories, as well as the category membership of individual elements, was not indicated to participants in any way and could not be deduced based on the surface visual properties of the stimuli. Category membership could be inferred only on the basis of five sources of information in the input to which learners may become sensitive during training. One is the serial order position of elements in strings. To our knowledge, this is the first experiment where SRT statistical learning research has used finite strings (i.e., in which stimuli are not presented as a continuous, unsegmented stream). If learners are differentially sensitive to elements depending on their serial order position, then RTs for a given category may differ depending on the location of that category in the string. One would expect to find primacy and recency effects, with categories at the beginnings and ends of strings showing preferentially faster RTs beyond the effects due to other sources of information. It should be the case that the effect of serial order position can be regressed out of the data so that the effects of the remaining four factors can be assessed.³ The remaining four factors are element probability, bigram RF, FCP, and BCP.

Initial analyses indicated that single element probability was not informative in and of itself, and it was therefore not analyzed further. In addition, bigram RF was equated for the transitions of interest during training. Thus, FCP and BCP were the only sources of information available to learners that could influence performance on any particular element transition. Of course, one could compute statistics across transitions of any length, or even across non-adjacent elements. However, because elements from the categories preceding category C and following category X are assigned at random and are independent of any specific CX element transition, the statistics associated with transitions other than from C to X are uninformative with regard to the ability to predict X. Hence, analyses are limited to adjacent transitions of length 2.

The statistics associated with each of the category transition subtypes during training are shown in the left half of Table 2. Previous research (Hunt & Aslin, 2001) indicates that participants are sensitive to both RF and FCP in continuous stimulus streams. The lower portion of Table 2 shows the statistics associated with the category transition subtypes during testing.

Procedure

Participants were seated alone in a private room. They were instructed to respond to the shapes that appeared in the stimulus box by tapping the matching shape in the response frame. No mention was made of RTs, learning, or of patterns embedded in the stimuli. Participants were instructed simply to tap the matching shape as rapidly as possible while maintaining high accuracy. Participants were told to respond only with their dominant hand and to use a single finger. Participants were monitored via closed circuit video to ensure compliance.

At the beginning of a string a green “start” circle appeared in the stimulus box and a second green circle appeared in the center of the response frame. When the participant tapped the green circle in the center of the response frame, both it and the green circle in the stimulus box immediately disappeared and a new shape, the first in the string, appeared in the stimulus box with no delay. When the participant tapped a shape in the response frame, whether it matched

³In both experiments, two regression analyses were conducted for each participant individually to factor out the effects of individual kinematics (i.e., idiosyncratically fortuitous or non-fortuitous assignment of grammatical elements to locations in the response grid) and of serial order position of elements within stimulus strings. For the kinematic regression, each correct RT was coded for the vector of the hand movement associated with that RT. This vector was coded using the following seven factors: vertical and horizontal grid location of the starting element, vertical and horizontal grid location of the target element, magnitude of the movement vector in grid units, and sine and cosine of the angle of movement. These factors were entered into a stepwise regression analysis with raw RT as the dependent measure. Residuals from this analysis were then used as the dependent measure in a serial order position regression. Serial order was predicted to produce slower RTs to elements in the middle of strings than to elements at the ends of strings. Accordingly, a quadratic curve was fit to the residuals from the kinematic regression and a second regression analysis was performed with kinematic residual RT as the dependent measure and the fit for serial order position as the independent variable. Residuals from this second regression analysis were then submitted to further analyses.

the shape in the stimulus box or not, the next shape in the string immediately appeared in the stimulus box. This process continued until the end of the string, at which point a red “stop” circle appeared in the stimulus box. No response was required for the red circle, which disappeared after .5 seconds. After a subsequent, randomly variable .5 to 1.0 second delay, the next string began. The variable delay prevented participants from accurately anticipating the next string onset. Participants could pause as long as they wished before tapping the green “start” circle (though long pauses were rare). Shapes in the response frame were always visible.

No feedback was given regarding correct or incorrect responses when the computer was able to localize a tap. However, for ambiguous responses involving taps on the border area between boxes, and for taps so light that a location could not be determined, the screen would flash red. The computer would then wait for the participant to provide an interpretable response before advancing to the next stimulus.

Participants took part in the experiment for 5 consecutive days for approximately 1 hour per day. On the first day, the 16 shapes that appeared in the response frame were selected randomly for each participant from a pool of 25 possible shapes. Additionally, the location of each shape in the frame was randomly assigned for each participant. Finally, the elements in the grammar were assigned randomly to locations in the response grid for each participant. These participant-unique randomizations were maintained throughout the experiment. The first 4 days of the experiment consisted of training blocks. Each training day had 7 blocks, with a 30-second rest period between blocks. Each block consisted of 49 strings drawn at random from the training set, for a total on each training day of 343 strings.

The fifth day consisted of 8 blocks. The first block was identical to the training blocks. It was included to place participants back in the same situation they had experienced on previous days before confronting them with strings that differed from those encountered during training. The remaining 7 blocks were test blocks comprised of 49 strings each, 28 drawn from the training set, 7 drawn from the generalization set, and 14 drawn from the illegal category extension set. All strings were drawn randomly from their respective sets. The average string length across all 5 days was 5.5 elements (50% length 5, 50% length 6).

Data collection—Grid location and RT data were collected for each response. RT was defined as the time from stimulus appearance to the first tap on the screen. A total of approximately 7546 measurements were possible across the 4 training days, with an additional 270 measurements possible during the first block of Day 5. Across the final 7 test blocks on Day 5, a total of approximately 1887 measurements were possible.

Results and Discussion

The bigram RFs and FCPs for the critical Generalization and Illegal category extension transitions are 0 when initially encountered. It is only with continued exposure on Day 5 that the statistics in the testing section of Table 2 accumulate. If participants are initially sensitive to the fact that they have never encountered the Generalization or Illegal transitions, then RTs to these transitions should be slower than RTs to either of the familiar transitions (Low Training and High Training). Additionally, if participants are responding solely on the basis of the available statistics, RTs to Illegal transitions should be no different than RTs to Generalization transitions. Thus, the key issue of whether learners show evidence of category generalization beyond the statistics of the input will be reflected in how RTs to Generalization transitions compare to RTs to Low and High Training transitions on the one hand, and Illegal transitions on the other.

If participants have learned that only certain elements can occur in the category X context, RTs to Generalization transitions should be faster than those to Illegal transitions. There are two

patterns of results that would be consistent with this. In the first, RTs to Generalization transitions are as fast as transitions encountered during training (Low and High Training). This would indicate that participants ignore the transitional statistics associated with Generalization transitions and do not differentiate them from transitions they saw during training. In the second pattern, RTs to Generalization transitions are slower than those to Low and High Training transitions, but not as slow as those to Illegal transitions. This would be the case if participants were sensitive to the fact that certain elements are licensed for the context of category X, while others were not, and at the same time remained sensitive to transitional information that allowed them to distinguish previously encountered transitions from novel transitions. Both of these patterns would be evidence of having formed a category. If participants have *not* generalized category membership, but instead treat all novel transitions the same, then a third pattern of results should occur. In this case, RTs to Generalization transitions are different from Low and High Training transitions, but are the same as RTs to Illegal transitions. This is what would be expected if learners were veridical in response to the available statistics during training.

RT data from Experiment 1 are presented in Figure 3 after regressing out the effects of kinematics and serial order position for each participant separately (all participants showed a significant effect of serial order position, $p < .05$; see footnote ³). Only correct responses were submitted to analysis. An overall learning effect is evident in the pattern of decreasing reaction times across sessions. Error rates are presented in Table 3.

To determine which of the three possible patterns of results was supported by the data, an ANOVA was conducted on the residual RTs of the four CX subtypes from the first three test blocks on Day 5 (blocks 30–32; test blocks contained generalization and illegal strings, in addition to training strings), with block and transition subtype as within-subjects factors.⁴ The analysis revealed a significant main effect of transition subset, $F(3, 57) = 11.782$, $\epsilon_{H-F} = .859$, $p < .001$, $MSE = .01518$, $\eta^2 = .383$.⁵ All other effects were non-significant. In order to assess whether the ANOVA findings may have masked an effect of serial order position, the data analysis was reformulated as a general linear regression model (GLM). This analysis was based on the residual RT scores that contributed to the means that were submitted to the ANOVA. Each correct residual RT from the first three test blocks on Day 5 was coded for block, transition subset, and serial order position. A GLM with block, transition subset, and serial order position as with-subjects factors revealed a significant main effect of transition subset ($F(3, 57) = 15.7092$, $p < .0001$, $MSE = .1198$, $\eta^2 = .453$). All other effects were non-significant, indicating that differences among CX subtypes are not due to an effect of serial order position.

Planned comparisons of the main effect of transition subset were conducted in order to fully evaluate the pattern of responses. 95% confidence intervals (CI) around the difference between means (M_d) for these comparisons are reported below. There is a minimum criterion that must be met for the overall pattern of data to be interpretable. Specifically, the RTs to transitions that involve ungrammatical category members (i.e., Illegal transitions) must be slower than reaction times to either of the two types of transitions encountered during training (i.e., Low and High Training transitions). Our data meet this criterion (Low Training vs. Illegal, $M_d = -.144$, $CI = -.156, -.071$; High Training vs. Illegal, $M_d = -.144$, $CI = -.174, -.054$). Another

⁴As noted elsewhere, the statistics of the input change on Day 5 as a consequence of the introduction of test strings. Because it was unclear how quickly learners' behavior would adapt to this change, and because we were interested in learners' initial response to these novel strings rather than adaptive changes in behavior as a consequence of continued exposure to them, we chose to analyze a subset of the test data from Day 5. At the same time, we needed to include enough data in our analysis to have sufficient power to detect any effects that might be present. Therefore, although data were collected from seven test blocks on Day 5, data were analyzed from the first three of those blocks (blocks 30–32).

⁵Throughout the paper, uncorrected degrees of freedom and MSE are reported for F tests. However, whenever a significant Mauchley's test ($p < .05$) indicates a violation of sphericity, the relevant epsilon (ϵ) correction factor is presented, and the associated p -value for the test reflects the adjusted degrees of freedom. The Greenhouse-Geisser epsilon (ϵ_{G-G}) is used when it is less than or equal to 0.75. In cases where ϵ_{G-G} exceeds 0.75, the Huynh-Feldt epsilon (ϵ_{H-F}) is used (see Keppel, 1991).

basic prediction is that RTs for the types of transitions encountered during training (i.e., Low and High Training) will not differ. The data support this prediction (Low vs. High Training, $M_d = .001$, $CI = -.035, .036$). The final predictions concern transitions that involve novel transitions between legal category members (i.e., Generalization transitions). The strongest evidence of having formed a category for the elements of X would be for RTs to Generalization transitions to be equivalent to RTs for Low and High Training transitions, but different from Illegal transitions. The data support these predictions as well (Low Training vs. Generalization, $M_d = -.017$, $CI = -.064, .029$; High Training vs. Generalization, $M_d = -.018$, $CI = -.059, .023$); Generalization vs. Illegal, $M_d = -.096$, $CI = -.149, -.043$).

The pattern of RT results from Experiment 1, in which Low Training, High Training, and Generalization transitions are not different from one another, but are each significantly faster than Illegal transitions, is the pattern most consistent with learners having induced a category for the elements in X. It indicates that all grammatical transitions, whether familiar or novel, are treated equivalently, but that transitions with category violations are not. Nevertheless, close inspection of the inset in Figure 3 indicates that responses to Generalization transitions are somewhat slower than they are to High Training transitions. Thus, although the pattern of RT results suggests that by the end of training, learners have induced a category for the members of X, there is also evidence of sensitivity to the difference between familiar grammatical strings and novel legal extensions, suggesting that category learning is not perfect and/or that within-category sensitivity is maintained. One possibility is that learners may have been sensitive to the exclusive relationships between elements in the High Training transitions. That is, the presence of a small number of perfectly predictable backwards transitions (i.e., element 4 was always and only preceded by element 9, and element 5 was always and only preceded by element 8) may have been particularly salient. If this were the case, it may have facilitated discrimination between strings with High Training transitions encountered during training and strings with novel, but legal, category extensions (i.e., Generalization transitions). To explore whether these findings were specific to the design of Experiment 1, we implemented a change in Experiment 2 to eliminate the exclusive relationships between elements in the High Training transitions.

Experiment 2

Experiment 2 sought to address the effect of context density by implementing a modest change in the available distributional statistics. We increased the contextual density of the X category by adding an additional element to category C. This increased from 1 to 2 the number of transitions to each X element in the High Training transitions. Transition restrictions into category X for this new element were similar to those for other elements of category C. We predicted that learners would again show evidence of having induced a category for the elements in X based on their generalization to legal, as opposed to illegal category extensions.

Method

Participants—Data were collected from 20 adults (9 men and 11 women; all right-handed) ages 18 to 33, all of whom were undergraduate or graduate students at the University of Rochester. Participants were paid \$90.00 for approximately 4.5 hours of participation over the course of 5 days, regardless of performance.

Apparatus—The apparatus was identical to that used in Experiment 1.

Design—The grammar used to generate strings in Experiment 2 is shown in Figure 4A. Strings were generated as in Experiment 1. This grammar differs from that in Experiment 1 by the addition of element 16 into category C. Element 16 was constrained in its transition to elements

in category X in that transitions to element 3 were prohibited during training. Training, generalization, and illegal category extension string sets were created as in Experiment 1.

The grammar in Figure 4A is capable of generating 360 unique string types. Of these, there are 288 possible training strings and 72 untrained generalization strings. Of the 288 possible training strings, 234 form the statistically balanced training set. The illegal extension set contains 72 string types. As in Experiment 1, the non-category B string types were doubled in each set. Thus, the training set contained 234 string types and 312 tokens, the legal extension set contained 72 types and 96 tokens, and the illegal extension set contained 72 types and 96 tokens.

As in Experiment 1, the constraints used to generate strings in Experiment 2 produced four subsets of transitions between elements in category C and elements in category X (i.e., Low Training, High Training, Generalization, and Illegal). All of the category and category transition subtypes for Experiment 2, along with their constituent elements and element transitions, are given in the right half of Table 1. The statistics associated with the strings in the training set, as well as those associated with testing, are shown in the right half of Table 2. Although this design changes the training statistics associated with several category transition subtypes from those present in Experiment 1, the most important are those that occur for Low and High Training transitions. Compared to Experiment 1, the bigram RFs for these transition types are at lower absolute values in Experiment 2, though the FCPs are the same. In addition, the BCPs for Low and High Training transitions are at lower absolute values than in Experiment 1, and the BCP for High Training transitions is no longer 1.0.

Procedure

The procedure was identical to that used in Experiment 1.

Results and Discussion

Predictions for the RT measure are based on the effects of introducing an additional element into category C. This addition has two consequences that affect the complexity of the information confronting participants. The first is that the total number of strings generated by the grammar increases by about 50%. This means that for any given number of training strings, exposure to the potential learning space in Experiment 2 will be sparser than in Experiment 1. In addition, the number of element transition types for CX transitions during training has increased from 8 in Experiment 1 to 12 in Experiment 2. Thus, any given element transition type is encountered fewer times over a given number of training strings in Experiment 2 than in Experiment 1. Intuitively, these two consequences should make it more difficult for learners to distinguish grammatical strings they encounter during training from strings that contain legal category extensions (generalizations). In contrast, the more extreme change introduced by a category violation should still be distinguishable from familiar grammatical strings.

RT data from Experiment 2 are presented in Figure 5 after regressing out the effects of kinematics and serial order position for each participant separately (all participants again showed a significant effect of serial order, $p < .05$). Only correct responses were submitted to analysis. An overall learning effect is evident in the pattern of decreasing reaction times across sessions. Error rates are presented in Table 3.

The three possible patterns of performance for the RT data are the same for Experiment 2 as for Experiment 1. To determine which pattern was supported by the data, an ANOVA was conducted on the residual RTs of the four CX subtypes from the first three test blocks on Day 5 (blocks 30–32) with block and transition subtype as within-subjects factors. A significant main effect of transition subset was revealed, $F(3, 57) = 19.031$, $\epsilon_{G-G} = .657$, $p < .001$, MSE

$= .01498, \eta^2 = .500$. All other effects were non-significant. As in Experiment 1, the data analysis was then reformulated as a GLM with block, transition subset, and serial order position as within-subjects factors. This revealed a significant main effect of transition subset ($F(3,57) = 15.8788, p < .0001, MSE = .1785, \eta^2 = .455$). All other effects were non-significant, indicating that, again, the differences among CX subtypes are not due to an effect of serial order position.

Planned comparisons of the main effect of transition subset were conducted in order to fully evaluate the pattern of responses. As in Experiment 1, the results from Experiment 2 show that responses to Illegal transitions are different from Low Training, High Training, and Generalization transitions (Low Training vs. Illegal, $M_d = -.159, CI = -.222, -.096$; High Training vs. Illegal, $M_d = -.128, CI = -.163, -.093$; Generalization vs. Illegal, $M_d = -.099, CI = -.146, -.052$). This indicates that transitions containing category violations (i.e., Illegal transitions) are treated differently from transitions that do not. In addition, as in Experiment 1, responses to Low Training transitions do not differ from High Training transitions (Low vs. High Training, $M_d = -.031, CI = -.085, .022$). This indicates that familiar transitions are treated the same whether they involve X1 or X2 category subtype elements. Performance for Generalization transitions is slower than for both Low and High Training transitions, but is significantly different only from Low Training transitions (Low Training vs. Generalization, $M_d = -.060, CI = -.101, -.019$; High Training vs. Generalization, $M_d = -.029, CI = -.063, .005$).

The pattern of responses for Low Training, High Training, and Generalization transitions suggests that learners treat transitions with familiar (Low Training and High Training) and novel (Generalization) legal transitions as containing legal category members, but still manage to discriminate among them. It is interesting that, after increasing the contextual density of the X category in Experiment 2, responses to Generalization transitions are somewhat slower than both Low Training and High Training transitions. This indicates that participants are sensitive to the fact that the Generalization transitions are different from those they encountered during training, even though they involve legal category members.

General Discussion

The results of both experiments are consistent with learners having induced a category for the elements that comprise X using only the information available in the distribution of those elements across a learning corpus. RT data from Experiment 1 also suggested that learners maintained the ability to distinguish among different legal transition types (Low Training, High Training, and Generalization transitions). In Experiment 2, the contextual density of the X category was increased by introducing an additional element into the category that precedes X. RT data from Experiment 2 indicate that learners maintained the ability to distinguish among the fine-grained differences in transitions between Low Training, High Training, and Generalization transitions, while still showing evidence of having induced a category for the elements in X.⁶

These results are reminiscent of findings from the categorical perception literature. Pisoni and Tash (1974) found that in a speeded same-different categorization task using pairs of speech sounds, participant responses were faster for pairs of acoustically identical stimuli than for pairs of acoustically different stimuli that were members of the same category. They interpreted this as evidence that participants have access to low-level acoustic information about speech stimuli along with more abstract phonetic representations. They further argued that access to

⁶Reaction time data indicate that learning occurred throughout the four days of training in each experiment. The number of participants and days of training were chosen based on pilot data from a similar paradigm. Although it may be the case that reliable evidence of category induction could be found with less training, we were not able to assess this because our test of category generalization occurred at the end of training.

low-level information depends in part on the type of information processing task employed. Other evidence for dual levels of processing, although not strictly categorical in nature, can be found in the literature on probabilistic phonotactics. Vitevitch and Luce (1998, 1999) showed that high probability phonotactic patterns were processed faster than patterns with low phonotactic probability, but only for stimuli that involved non-words. Speed of processing for words was determined by neighborhood density. They argue that probabilistic phonotactics dominates performance at the sublexical level of processing, but that competition among lexical neighbors dominates when test items involve words. Similar level-of-processing arguments can be employed in interpreting the results from the current experiments. That is, over the course of training, participants not only develop and have access to categorical distinctions among elements, but also develop and have access to distinctions among transitions between elements.

In contrast, Perlman, Pothos, Edwards, and Tzelgov (in press) have demonstrated that certain forms of learning (i.e., chunking) in the SRT paradigm can be driven by task demands rather than available co-occurrence statistics. It is likely that in real-world learning, a wide variety of computational and contextual influences (exogenous and endogenous) impact the nature of what is learned. In the present study, task demands were constant across experimental conditions and thus were not a factor in category formation. It seems likely, however, that task demands (or other factors) might be capable of either reinforcing the category induction process demonstrated here or diluting it. An interesting, but unanswered question is whether participants might show behavioral evidence of having become sensitive to co-occurrence statistics even in a situation where the measure of learning is driven by other factors, such as task demands. Given findings from the speech perception literature as well as our own work in statistical learning, we expect that evidence for such sensitivity could be found, though the effects might be subtle.

One critique of the current experiments is that an alternative hypothesis may partly account for the differences in RT between legal (i.e., training and generalization) transitions and illegal transitions. Over the course of training, participants may learn the association between a specific sequence position and a set of likely responses. Although we attempted to reduce the effect of this response-probability-by-position association by using strings of varying length, we did not eliminate it. Thus, the difference in RT between legal and illegal transitions may be due, in part, to a violation of this probability-by-position association in the case of illegal transitions. The current data can not address the potential contribution of this effect to our findings. However, this type of association likely does not account for all of the observed differences between legal and illegal transitions in our data because we also demonstrate differences between training transitions and generalization transitions in Experiment 2. In particular, the difference in RT between low-training and generalization transitions cannot be due to differences in learned probability-by-position associations, but instead must be based on sensitivity to the transitional statistics of the training input. To completely resolve the effect of response-probability-by-position would require an experiment in which two different categories occurred in the same sequential position, but were defined using unique surrounding contexts (as in linguistic sub-categorization). Illegal transitions could then be created by interchanging the elements of the two categories, without also changing their sequential position. Such an experiment may form the basis of future work.

There are two caveats to the current experiments in light of the debate on linguistic form-class induction outlined in the introduction. The first is that the experiments were conducted using adult participants, and hence do not directly address the ability of young learners to form categories based on distributional information. It should be noted that Mintz (1996) conducted an experiment in which he found that 9-month-old infants showed sensitivity to the ordering of nonsense words in sequences after a short period of training. In addition, Mintz (2006)

presented evidence that 12-month-old infants can use distributional analysis to successfully categorize novel nonsense verbs. However, as the author points out, because the experiment used natural language contexts, it is possible that learners relied on pre-existing knowledge of English grammatical structure to facilitate their analyses. Nevertheless, these studies are promising because they suggest that infants are sensitive to some aspect(s) of the distributional structure of the input they receive from the world around them. This is supported by findings from other work (Gerken, 2006; Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998) showing that infants are sensitive to distributional cues. Further experiments that capitalize on the types of designs used here might enable one to determine whether infants can in fact form distributionally defined categories.

The second caveat is that the current experiments address only some of the performance criteria that might be required to conclude that a form-class category had been induced. The current designs test two basic criteria for having formed a category and provide strong evidence that these criteria are both met. One is whether items for which the learner has evidence of membership in another category (i.e., Illegal transitions) are rejected as potential members of the category under study. The other is whether novel category transitions that do not violate category membership (i.e., Generalization transitions) are treated as similar to familiar transitions involving category members (i.e., Low and High Training transitions). Two additional, more stringent criteria have not been addressed in the current experiments. The first criterion is whether elements with which learners are familiar, but which occur in only one type of grammatical context, will be generalized to other grammatical contexts for which that element should be a legitimate category member. An example from English would be the ability of a native speaker to generalize the use of a particular noun learned *only* as the object of a preposition to other grammatical contexts in which nouns are licensed, such as the complement of a determiner. The second criterion concerns whether a completely novel category exemplar introduced in one grammatical context will be generalized immediately to other contexts that are also licensed for that form-class. An example would be the ability to generalize a single instance of a novel nonsense word introduced as the object of a preposition to another grammatical context in which nouns are licensed, such as the complement of a determiner.

Although the current experiments do not address these last two criteria, the evidence presented indicates that learners can utilize transitional information not only in service of category formation, but also to discriminate among within-category members under specifiable conditions. Thus, although the form-class membership of a particular element may be categorical, there are subtle behavioral differences among legitimate category members that are contingent on the evidence learners have accumulated over time. This, perhaps, is not surprising. What is interesting, however, is that we now have evidence that the pair-wise transitional statistics of relative frequency, forward conditional probability, and backward conditional probability are utilized in the service of these behaviors, and that the ability to discriminate among category sub-types depends in part on the structure of the statistical evidence presented to learners. Evidence for this type of within-category sensitivity in a category that was induced as a function of specific transitional statistics has not been demonstrated in previous work that has used end-point grammaticality judgments of category induction (e.g., Mintz, 2002). The added theoretical advance here is that not only can learners induce categories based on transitional statistics, but that the structure of those statistics matters and can produce differential results depending on the specifics of that structure.

This is important because it provides empirical evidence of a mechanism that may be relevant for the induction of form-class categories. Note, however, that we are not claiming that distributional analysis is sufficient, in and of itself, for this purpose. There seems little doubt that other sources of information (among them semantics, prosody, and morphology) may also be relevant to the task of form-class induction. We are also not claiming that the role of

distributional analysis is primary to other types of analysis, such as semantic analysis. We are agnostic with regard to competing empiricist claims regarding form-class development (i.e., the semantic analysis of Bates & MacWhinney [1979, 1982] and the distributional analysis of Maratsos & Chalkley [1980]). The fact that there is sufficient distributional structure in the speech to young children for a learner to induce, in principle, basic categories (Mintz, 1996; Mintz, Newport, & Bever, 2002), and the evidence from the current experiments and in Mintz (2002) that adult learners can utilize this type of information to induce categories, together provide promising support for the hypothesis that distributional analysis may play a significant role in form-class induction during the process of natural language acquisition. The work presented here is one step in the direction of refining the question of what aspects of language acquisition may be, at least in principle, a consequence of domain-general computational mechanisms, and characterizing the kinds of statistics over which these mechanisms operate.

Acknowledgments

We would like to thank Elissa L. Newport for productive conversations at many points in the conceptualization and planning of this work. We would also like to thank Jennifer Hooker, Elizabeth Gramzow, and Kileen McCrink for their assistance in data collection. This research was conducted as part of the first author's doctoral dissertation at the University of Rochester, and was supported in part by funding from the NSF (SBR-9873477), the Packard Foundation (2001-17783), and the ONR (N00014-07-1-0937).

References

- Aslin RN, Saffran JR, Newport EL. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 1998;9:321–324.
- Bates, E.; MacWhinney, B. The functionalist approach to the acquisition of grammar. In: Ochs, E.; Schieffelin, B., editors. *Developmental pragmatics*. New York, NY: Academic Press; 1979. p. 167-211.
- Bates, E.; MacWhinney, B. Functionalist approaches to grammar. In: Wanner, E.; Gleitman, LR., editors. *Language acquisition: The state of the state of the art*. Cambridge: Cambridge University Press; 1982. p. 173-218.
- Braine MDS. On learning the grammatical order of words. *Psychological Review* 1963;70:323–348. [PubMed: 14014895]
- Brooks LR, Vokey JR. Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: General* 1991;120:316–323.
- Chomsky, N. *The logical structure of linguistic theory*. New York: Plenum Press; 1955/1975.
- Chomsky, N. *Syntactic structures*. Mouton: The Hague; 1957.
- Cleeremans A, McClelland JL. Learning the structure of event sequences. *Journal of Experimental Psychology: General* 1991;120:235–253. [PubMed: 1836490]
- Cohen A, Ivry RI, Keele SW. Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 1990;16:17–30.
- Dienes, Z.; Altman, G. Transfer of implicit knowledge across domains: How implicit and how abstract?. In: Berry, D., editor. *How implicit is implicit learning?*. Oxford: Oxford University Press; 1997. p. 107-123.
- Fiser J, Aslin RN. Statistical learning of higher order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2002;28:458–467.
- Gerken LA. Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition* 2006;98:B67–B74. [PubMed: 15992791]
- Gomez RL, Gerken L. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 1999;70:109–135. [PubMed: 10349760]
- Gomez RL, Gerken L, Schvaneveldt RW. The basis of transfer in artificial grammar learning. *Memory and Cognition* 2000;28:253–263.

- Gomez RL, Schvaneveldt RW. What is learned from artificial grammars? Transfer tests of simple association. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 1994;20:396–410.
- Hasher L, Zacks RT. Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist* 1984;39:1372–1388. [PubMed: 6395744]
- Hunt RH, Aslin RN. Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General* 2001;130:658–680. [PubMed: 11757874]
- Keppel, G. Design and analysis: A researcher's handbook. 3rd ed.. Upper Saddle River, NJ: Prentice-Hall; 1991.
- Kiss GR. Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation* 1973;7:1–41.
- Lashley, KS. The problem of serial order in behavior. In: Jeffress, LA., editor. *Cerebral mechanisms in behavior: The Hixon symposium*. New York: Wiley; 1951. p. 112-136.
- Maratsos, M.; Chalkley, MA. The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In: Nelson, K., editor. *Children's language*. Vol. 2. New York: Gardner Press; 1980. p. 127-189.
- Marcus GF, Vijayan S, Bandi Rao S, Vishton PM. Rule learning by seven-month-old infants. *Science* 1999;283:77–80. [PubMed: 9872745]
- Mintz, TH. The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories. University of Rochester; 1996. Unpublished Doctoral Dissertation
- Mintz TH. Category induction from distributional cues in an artificial language. *Memory & Cognition* 2002;30:678–686.
- Mintz TH. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 2003;90:91–117. [PubMed: 14597271]
- Mintz, TH. Finding the verbs: Distributional cues to categories available to young learners. In: Hirsch-Pasek, K.; Golinkoff, RM., editors. *Action meets word: How children learn verbs*. New York: OUP; 2006. p. 31-63.
- Mintz TH, Newport EL, Bever TG. The distributional structure of grammatical categories in speech to young children. *Cognitive Science* 2002;26:393–425.
- Morgan JL, Meier RP, Newport EL. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology* 1987;19:498–550. [PubMed: 3677585]
- Monaghan P, Chater N, Christiansen MH. The differential role of phonological and distributional cues in grammatical categorization. *Cognition* 2005;96:143–182. [PubMed: 15925574]
- Nissen MJ, Bullemer P. Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology* 1987;19:1–32.
- Perlman A, Pothos EM, Edwards DJ, Tzelgov J. Task-relevant chunking in sequence learning. *Journal of Experimental Psychology: Human Perception and Performance*. in press.
- Pisoni DB, Tash J. Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics* 1974;15:285–290.
- Pothos EM. Theories of artificial grammar learning. *Psychological Bulletin* 2007;133:227–244. [PubMed: 17338598]
- Radford, A. *Transformational grammar: A first course*. Cambridge: Cambridge University Press; 1988.
- Redington M, Chater N, Finch S. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 1998;22:435–469.
- Reber AS. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 1967;6:855–863.
- Reber AS. Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology* 1969;81:115–119.
- Reber AS. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 1989;118:219–235.

- Reber, AS. Implicit learning and tacit knowledge: An essay on the cognitive unconscious. New York: Oxford University Press; 1993.
- Reed J, Johnson P. Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 1994;20:585–594.
- Reeder, PA.; Newport, EL.; Aslin, RN. The role of distributional information in linguistic category formation. Paper presentation at the Cognitive Science Society meeting; Amsterdam. 2009 Aug.
- Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts; 1972. p. 64-99.
- Saffran JR. The use of predictive dependencies in language learning. *Journal of Memory and Language* 2001;44:493–515.
- Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. *Science* 1996;274:1926–1928. [PubMed: 8943209]
- Saffran JR, Newport EL, Aslin RN. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 1996;35:606–621.
- Stadler MA. Statistical structure and implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 1992;18:318–327.
- Vitevitch MS, Luce PA. When words compete: Levels of processing in spoken word recognition. *Psychological Science* 1998;9:325–329.
- Vitevitch MS, Luce PA. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 1999;40:374–408.

A

B

Figure 1.

(A) Schematic of the visual display seen by right-handed participants. (B) Sample sequence of shapes that appeared one-at-a-time in the stimulus box, beginning with a green “start” circle (indicated by the “G”) and ending with a red “stop” circle (indicated by the “R”).

A

A	(B)	C	X	D	E
6	14	8	1	10	12
7	15	9	2	11	13
			3		
			4		
			5		

B

A	(B)	C	X	D	E
6	14	8	13	10	12
7	15	9	12	11	13

Figure 2.

(A) The grammar used in Experiment 1. Letters refer to categories. Numbers under a given letter refer to the elemental members of that category, which corresponded to specific shapes in particular spatial locations, assigned uniquely for each participant. Dashed lines between numbers indicate transitions which did not occur during training, but which did occur during testing as instances of legal category extensions (Generalization transitions). Parentheses around category B indicate that it occurred optionally, 50% of the time. (B) Alteration to the grammar used in Experiment 1 to generate strings with illegal category extensions (Illegal transitions). Note that for any given string, the element that occurred in category X could not be repeated in category E. Dashed lines between numbers indicate these prohibited transitions.

Figure 3. Mean residual RTs for each category transition subtype in Experiment 1 (N=20). Residuals are the result of a regression analysis that evaluated the effects individual kinematics and serial order position on RT (see footnote ³). CX category transition subtypes used in statistical analyses are highlighted in the inset.

A

B

Figure 4.

(A) Grammar used in Experiment 2. Parentheses around category B indicate that it occurred optionally, 50% of the time. (B) Alteration to the grammar used in Experiment 2 to generate strings with illegal category extensions (Illegal transitions).

Figure 5. Mean residual RTs for each category transition subtype in Experiment 2 (N=20). Residuals are the result of a regression analysis that evaluated the effects individual kinematics and serial order position on RT (see footnote ³). CX category transition subtypes used in statistical analyses are highlighted in the inset.

Table 1

Categories, elements, and element transitions comprising the different category and category transition subtypes in Experiments 1 and 2. Category and category transition subtypes are comprised of elements and element transitions, respectively. Note that G represents the green “start” circle at the beginning of every stimulus string.

Elements		
Category	Experiment	Experiment 2
G	G	G
A	6 7	6 7
B	14 15	14 15
C	8 9	8 9 16
X1	1 2 3	1 2
X2	4 5	3 4 5
X3	4 5	3 4 5
X4	12 13	12 13
D	10 11	10 11
E	12 13	12 13
Element Transitions		
Category Transition	Experiment 1	Experiment 2
GA	G-6 G-7	G-6 G-7
AB	6-14 6-15	6-14 6-15
BC	14-8 14-9	14-8 14-9
AC	6-8 6-9	6-8 6-9
CX1 Low Training	8-1 8-2	8-1 8-2
CX2 High Training	8-4 9-5	8-3 8-4
CX3 Generalization	8-5 9-4	8-5 9-4
CX4 Illegal	8-12 8-13	8-12 8-13
X1D	1-10 1-11	1-10 1-11
X2D ^a	4-10 4-11	3-10 3-11
	5-10 5-11	4-10 4-11
	7-14 7-15	7-14 7-15
	15-8 15-9	14-16 15-8
	7-8 7-9	6-16 7-8
	8-3 9-1	9-1 9-2
	9-2 9-3	8-2 9-3
	9-13 9-12	16-3 16-12
	2-10 2-11	9-12 9-13
	3-10 3-11	2-10 2-11
	5-10 5-11	4-10 4-11
	7-14 7-15	15-8 15-9
	15-8 15-9	7-8 7-9
	8-3 9-1	16-1 16-2
	9-2 9-3	16-4 16-5
	9-4 9-5	16-3 16-12
	8-12 8-13	16-12 16-13
	1-10 1-11	2-10 2-11
	4-10 4-11	5-10 5-11

Elements									
Category	Experiment					Experiment 2			
X4D	12-10	12-11	13-10	13-11	12-10	12-11	13-10	13-11	
DE	10-12	10-13	11-12	11-13	10-12	10-13	11-12	11-13	

^aTransitions from X2 and X3 to D can not be distinguished based solely on element pairs, and have been collapsed into X2D.

Table 2

Statistics associated with the stimulus sets in Experiments 1 and 2. Three statistics are presented for a given transition between two sample elements X and Y belonging to two categories: bigram relative frequency, forward conditional probability, and backward conditional probability. The values presented reflect the mean, across participants, of statistics computed from the actual stimulus sets presented to individual participants. Values for the training set reflect statistics computed through the end of Block 29. Values for the testing set reflect statistics computed through the end of Block 32, which was the last block for which RTs were submitted to statistical analysis. The heavy boxes highlight the critical experimental transitions.

Training (Blocks 1–29)						
Experiment 1				Experiment 2		
Category Transition	RF	FCP	BCP	RF	FCP	BCP
GA	0.0909	0.500	1.000	0.0909	0.500	1.000
AB	0.0226	0.249	0.500	0.0227	0.249	0.500
BC	0.0226	0.500	0.249	0.0151	0.333	0.249
AC	0.0228	0.251	0.251	0.0152	0.167	0.251
CX1 Low Training	0.0227	0.250	0.500	0.0151	0.249	0.333
CX2 High Training	0.0227	0.250	1.000	0.0152	0.251	0.500
X1D	0.0227	0.500	0.250	0.0227	0.500	0.249
X2D	0.0114	0.500	0.125	0.0152	0.500	0.167
DE	0.0455	0.500	0.500	0.0455	0.500	0.500
Testing (Blocks 1–32)						
Experiment 1				Experiment 2		
Category Transition	RF	FCP	BCP	RF	FCP	BCP
GA	0.0909	0.500	1.000	0.0909	0.500	1.000
AB	0.0228	0.251	0.500	0.0227	0.250	0.500
BC	0.0228	0.500	0.251	0.0151	0.333	0.250
AC	0.0227	0.249	0.249	0.0152	0.167	0.250

Training (Blocks 1–29)						
Category Transition	Experiment 1			Experiment 2		
	RF	FCP	BCP	RF	FCP	BCP
CX1 Low Training	0.0218	0.240	0.500	0.0145	0.240	0.333
CX2 High Training	0.0218	0.240	0.947	0.0146	0.241	0.486
CX3 Generalization	0.0012	0.013	0.053	0.0008	0.013	0.027
CX4 Illegal	0.0012	0.013	0.013	0.0008	0.013	0.009
X1D	0.0218	0.500	0.240	0.0218	0.500	0.240
X2D	0.0115	0.500	0.127	0.0150	0.500	0.165
X4D	0.0012	0.013	0.013	0.0012	0.013	0.013
DE	0.0454	0.500	0.487	0.0455	0.500	0.487

Table 3

Average error rates per block for Experiments 1 and 2 (mean percentage and range).

Error Type	Experiment 1		Experiment 2	
	Training	Test	Training	Test
Incorrect Response	1.9 (1.3–3.2)	1.9 (0.9–3.9)	1.7 (0.8–2.9)	1.9 (0.7–3.7)
Border	1.9 (0.8–3.9)	1.9 (0.8–5.5)	1.6 (0.6–3.3)	1.6 (0.6–5.1)
Localization Difficulty	3.1 (1.3–7.2)	3.0 (0.7–7.3)	3.2 (1.2–8.3)	2.7 (1.0–7.6)
Total	6.1 (3.2–14.4)	6.0 (2.9–13.4)	5.9 (2.9–10.7)	5.6 (2.5–13.2)